

WORKING PAPER · NO. 2018-75

Examiner Inconsistency: Evidence from Refugee Appeals

Samuel Norris

MAY 2019

Examiner Inconsistency: Evidence from Refugee Appeals*

Samuel Norris[†]

May 1, 2019

Abstract

Different judges, doctors, loan officers, and patent examiners make different decisions, generating costly uncertainty over ultimate outcomes. In this paper, I use multiple-stage decision-making institutions to identify nonparametric bounds on disagreement between decision-makers. I bound disagreement to at least 17% of all Canadian refugee appeals, 150% larger than the estimate using existing methods and substantial relative to an average approval rate of 14%. I aggregate disagreement into judge-specific measures of quality, and find that quality improves with experience, declines with workload, and is higher for judges appointed under a nonpartisan regime. Finally, I adopt my method to test and reject the typical examiner-assignment monotonicity assumption.

*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, Kerwin Charles, Michael Frakes, Ezra Friedman, Jeff Grogger, 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, University of Chicago 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 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 for a very productive semester.

[†]Harris School of Public Policy, University of Chicago. samnorris@uchicago.edu

Many important economic decisions are made by individuals exercising their discretion on behalf of larger institutions, such as loan officers allocating credit and examiners awarding patents. Discretion lets institutions apply general principles to complicated, novel situations. However, since different decision-makers might decide the same case differently, uncertainty over the identity of the decision-maker generates uncertainty over the eventual outcome.

The costs of uncertainty are high. Uncertainty that affects future prices or costs—such as over litigation or whether a patent will be granted—distorts investment (Bernanke, 1983) and reduces welfare (Craswell and Calfee, 1986). Reflecting this, surveys use legal predictability as a measure of business friendliness (Djankov et al., 2003). For individuals, time out of the workforce awaiting the result of an uncertain disability insurance application erodes human capital (Autor et al., 2015).

This paper studies the degree and causes of inconsistency in institutions where multiple examiners classify cases into two groups—for example, guilty and not guilty—according to a common standard. This represents many important decisions, including 630,000 annual patent applications, 1.7 million cancer diagnoses, 2 million SSDI applications, and 47 million state court cases. I focus on decision processes with multiple rounds—for example, appeals or second medical opinions—and show that access to information on a second round allows much more accurate measurement of the prevalence with which two decision-makers would disagree on the correct decision. I first construct nonparametric bounds on disagreement, then build a structural model to understand the causes of inconsistency and evaluate the success of reforms aimed at improving decision quality. Unlike previous work, the methods in this paper have implications for the monotonicity assumption in examiner-assignment IV designs, and I provide new tools to directly test the hypothesis.

The empirical challenge is that because we do not typically observe the same decision made by multiple examiners, disagreement is never directly observed. Previous work has approached this issue in two ways. First, a large literature uses the leniency of randomly-assigned examiners as an instrument for the decision (Kling, 2006). This sheds light on inconsistency because under the usual IV assumptions, the complier share between two judges—equal to the difference in their incarceration rates—is precisely the share of cases on which they disagree. However, recent work has questioned the monotonicity assumption in examiner designs (Mueller-Smith, 2015). Without this assumption, the cross-judge difference in incarceration rates does not identify the level of disagreement, but provides a relatively uninformative lower bound. Judges who each incarcerate half of defendants might agree on all potential decisions, or none of them.

Second, a large literature has studied non-relevant factors—such as the outcome of a football game—that nonetheless affect decisions.¹ While these studies show that examiners are not always consistent with themselves, the effects are small relative to cross-examiner variation in decisions.

¹Eren and Mocan (2016) find that college football losses affect decisions by judges who attended that school. Judge behavior is also affected by TV news (Philippe and Ouss, 2016), the previous decision (Chen et al., 2016), the defendant’s birthday (Chen and Philippe, 2018), and the outdoor temperature (Heyes and Saberian, 2019).

In the first part of the paper I build on the examiner IV bound on first-round disagreement. When cases go through multiple rounds of decision-making, I show that decisions by the second-round examiners can be used as a type of covariate, allowing substantial tightening of the bounds. Intuitively, if two equally-severe first-round examiners would never disagree—in other words, they would approve the exact same cases—then the cases they send to the second round should have the same average outcome no matter which second-round examiner makes the ultimate decision. Deviation from this ideal can be translated into bounds on disagreement for each of the $J(J-1)/2$ unique pairs of judges. The information gained from the second round is substantial; in my context it increases the disagreement bound by 150% relative to the first-round-only bound.

I then turn to understanding the determinants of consistency. Examiners can disagree with each other for two reasons. First, examiners could have different standards, which leads to differential approval rates across examiners. In most courts there is substantial cross-judge variation in propensity to incarcerate: in Ohio, [Norris, Pecenco, and Weaver \(2018\)](#) find that even when random assignment makes the distribution of cases the same for all judges, incarceration rates increase by nearly 19 pp when moving from the 5th to the 95th percentile of judge severity.

Second, examiners’ ordering of the cases by strength could differ, and this ultimately affects disagreement. Differences in ordering cause disagreement even between equally-severe examiners; ranking a case 10th versus 11th matters when the examiners each approve only their 10 strongest cases.

I build a simple model that summarizes examiner behavior along these two dimensions. The model is a generalization of standard index models of choice, where examiners perfectly observe the scalar strength of each claimant’s case and approve them if it is larger than some examiner-specific threshold. In my model, examiners observe case quality with error and so rank the relative strengths of cases differently; the average size of this error (and the corresponding difference from the consensus ranking) is inaccuracy. Examiner behavior is thus collapsed from the $J(J-1)/2$ measures of pairwise disagreement into an examiner-specific standard, and examiner-specific accuracy. As in much theoretical work on decision-making (e.g., [Sah and Stiglitz 1986](#)), the probability of approval is increasing in case quality for all judges, but more accurate judges approve more high-quality cases. The two-parameter parsimony allows me to study examiner-specific factors that affect consistency, such as the level of experience and the method with which they were selected.

The model is identified in multiple-stage decision-making institutions where first round examiners are instructed to approve cases that are likely to be approved by a second-round examiner. This means that the success of first-round examiners at selecting cases that are subsequently approved is a measure of their adherence to collective institutional standards, and can tell us which of two pairwise inconsistent examiners is more accurate.² The same conditions used to generate a lower

²Two 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

bound on disagreement—random assignment of examiners to cases, and two examiners making consecutive decisions on each case—play a key role in identification. With the addition of regressors that affect examiner leniency but not errors, the model is nonparametrically identified and can be tractably estimated under parametric restrictions. It is applicable to institutions where examiners make similar decisions in different rounds, in the sense that the same case characteristics improve the likelihood of both first- and second-round approval.³ This approach is particularly useful when the institutional objective is ill-defined. Researchers cannot hope to measure whether an invention meets the criteria for patentability, or a defendant is actually guilty, and so evaluating institutions by the number of Type I or Type II errors is conceptually infeasible. In these situations, disagreement and accuracy are useful alternative measures of examiner and institutional quality.

I use these tools to study the 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 status by administrative decision-makers, 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 . . . 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, which make up about 70% of their caseload. Nonetheless, I find low levels of agreement between judges on which claimants’ appeals should be granted, corresponding to a meaningful impact on outcomes. If judges made no ordering errors—corresponding to the naive examiner-assignment measure of disagreement—differential leniency across first-round judges means that the average pair of judges would disagree on the correct decision for 7% of cases, which is already relatively large given that the judges approve only 14% of cases. I show that this severely understates the level of inconsistency, and bound disagreement to at least 17% of all cases.

The lack of consistency also means that the first-round judges fail to approve many claimants who could be successful in the second round. Using the structural model, I find that if all claimants automatically advanced to the second round, 19.4% of appeals would be successful, rather than the current 6%. This difference amounts to approximately 10,400 families over my study period.

The structural model uncovers several important predictors of decision quality, which could be used to improve the quality of decisions. First, judge accuracy improves dramatically during the first year of experience, and continues to improve at a slower rate for at least ten years. This

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.

³My setting is particularly attractive because the second-round decision occurs automatically. When the claimant must decide whether to appeal, this decision must be specifically accounted for. In terms of the bounding estimator, differences across first-round judges in appeal rates (or joint rates of appeal and appeal success under a specific second-round judge) speaks to cross-judge disagreement in the same way that success in the second round does in my context. In the structural model, the appeal decision can be modeled as an additional stage of decision-making.

suggests that the court could improve overall consistency by increasing judge retention.

Second, judges get better over time at maintaining consistency in periods of high workload; judges with fewer than five years of experience become less accurate with higher workloads, but experienced judges are unaffected by the number of cases. As a result, consistency could be further improved by transferring cases from inexperienced to experienced judges.

Third, the method of judge selection matters. In 1988, a new law made it more difficult for the government to appoint unqualified judges by requiring that all candidates be approved by an independent committee of legal experts. The reform had the intended effect of reducing the number of newly-appointed judges with ties to the party in power, and dramatically improved judge accuracy and decision quality.

Fourth, court administrators could directly use their knowledge of judge behavior to optimize judge assignment. Knowledge of judge quality appears to be widespread; I survey refugee lawyers and find that survey responses are correlated with my measures of judge accuracy. Using the model estimates to optimally assign judges to rounds, I find that the Federal Court could reduce its workload by approximately 18% while approving the same number of similar-quality claimants, saving \$4.6 million in judge salaries alone over the study period.

Finally, cross-judge disagreement on case ordering typically implies violations of the monotonicity condition required in examiner-assignment designs. In the Appendix I introduce new tests for monotonicity based on the methods from this paper. Recent work has shown that the necessary strength of the monotonicity assumption depends on whether the estimand is a LATE or an MTE. My tests accommodate both assumptions, and I show that at least in this setting, LATE monotonicity looks to be satisfied while MTE monotonicity is decisively rejected. In simulations, the resulting MTE bias is relatively large.

The paper proceeds in five parts. [Section 1](#) briefly discusses the institutional background. I bound examiner-pair disagreement in [Section 2](#). [Section 3](#) introduces a structural model that aggregates examiner-pair disagreement rates into a measure of examiner quality. [Section 4](#) contains the results, and [Section 5](#) concludes.

1 Institutional background and data

Initial refugee decisions in Canada are made by the Immigration and Refugee Board (IRB). The IRB is not amenable to analysis because the procedure to assign examiners to cases is opaque and non-random. My entire analysis therefore concerns the Federal Court, which hears appeals of IRB decisions.⁴ However, I begin by describing the IRB in enough detail to characterize the denied

⁴The 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.

claimants who appeal to the Federal Court. I then describe the Federal Court and the relevant institutional background.

1.1 Immigration and Refugee Board

Initial decisions on refugee claims are made by Members of the IRB, who evaluate whether the claimant has a “well-founded fear of persecution for reasons of race, religion, nationality, membership in a particular social group or political opinion” (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. Between-Member approval rates vary dramatically, from 16% at the 10th percentile to 82% at the 90th percentile. The non-random assignment of cases means that this difference might merely reflect cross-Member variation in strength of case rather than variation in severity, though the scope of the variation seems at odds with the possible extent of specialization (Rehaag, 2007). 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 appeal to the Federal Court. Approximately 65% of denied claimants file an appeal, which allows most claimants to stay in Canada until the Federal Court makes its final decision.⁵

IRB procedures for refugee decisions were broadly consistent between 1995 and 2012 (Grant and Rehaag, 2015). The only major policy change occurred in 2002. Before this, cases were heard by a two-Member panel unless the claimant agreed otherwise, and refugee status was granted if either Member recommended it. After the implementation of the Immigration and Refugee Protection Act (IRPA), cases were decided by a single Member. This policy change may have affected the distribution of case quality for the rejected claimants who appeal to the Federal Court. To allow for this possibility, in the analysis I allow the distribution of case quality to vary before and after IRPA.

1.2 Federal Court responsibilities and protocol

The Federal Court has jurisdiction over certain issues related to the federal government, with about 70% of their caseload devoted to refugee appeals. The scope of these appeals is limited. Under Canadian law, judges must show deference to administrative decisions, and so decide only whether the decision was “reasonable,” not whether it was “correct” (Rehaag, 2012).⁶

⁵The IRB occasionally rules that an application was “without merit,” in which case removal can be immediate.

⁶An unreasonable decision is 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.” This 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

Success at the Federal Court requires approval by two consecutive quasi-randomly assigned judges. First-round judges decide whether a claimant has an “arguable case” to make to a yet-to-be-determined second-round judge, who decides whether the original government decision was reasonable. A natural implication is that a good measure of first-round judge quality is their ability to find as many claimants as possible who will be successful in the second round, conditional on their overall approval rate.

The first-round judge makes her decision after reviewing written records from the IRB and briefs written by the lawyers for each side. If the judge decides against the claimant, the claim is rejected and they are usually deported. If they decide for the claimant, the case goes to the second round. Regardless of their decision, the first-round judge does not provide a written justification. This makes it relatively difficult for first-round judges to learn how their colleagues make decisions, and may contribute to high levels of disagreement.

During the second-round hearing, the judge questions the lawyers about their briefs and the IRB records, but very rarely reviews new evidence. The name of the first-round judge is not immediately available, and the judges typically don’t obtain this information. To reflect this, I model the second-round judge as unaware of the identity of the first-round judge—the first-round judge affects the second-round decision through her choice of which claimants to approve, but not through an information channel. This eliminates the possibility of multiple equilibria.

If a claimant is successful in the second stage, their case is usually sent back to the IRB for a new decision, but sometimes they are automatically granted refugee status. I will ignore this distinction in the empirical analysis.

Judge assignment works similarly in both stages. For the first round, judges are assigned to cases using a pre-set schedule; in each office the judges rotate through “leave duty.” The leave duty judge receives the cases on one day (usually a Monday), and is responsible for disposing of all of them. There is no review of the cases before they are assigned, and the leave duty schedule is not public. Previous research claims that this assignment is as good as random (Rehaag, 2007), and in Section 4.2 I show that judge leniency is uncorrelated with case and claimant characteristics. In the second round, cases are divided between judges without review of the contents and a computer program slots the cases into available hearing times.

1.3 Reform to selection of Federal Court justices

Federal Court justices are appointed by the Minister of Justice. Before a 1988 reform, the Minister had enormous discretion over appointments, and used them in part to reward supporters of his party (McKelvey, 1985). The reform aimed to curtail this behavior by creating independent committees

is in danger in his own country—is not allowed, while evidence about how the decision was made—for example, that the IRB Member had made a racially prejudiced statement—would typically be accepted.

to evaluate applicants in terms of competence, fairness, and ethical standards (Hausegger et al., 2010). Only candidates who met these standards could be approved by the Minister. By design, four of seven members were from the broader legal community and not directly appointed by the Minister, making it difficult to push through unqualified nominees.

The standards had bite—only about 40% of candidates were approved—and seem to have reduced the level of patronage. Before the reform, at least 47% of appointed judges had some involvement with the ruling party, ranging from financial contributions to running for office (Russell and Ziegel, 1991). Conversely, only 30% of post-reform judges donated to the party in power in the five years before their appointment (Hausegger et al., 2010), suggesting the new system reduced the number of unqualified judges and improved the overall quality of the courts. I will test this hypothesis in Section 4.9, and find evidence that the reform improved consistency.

2 Bounding judge disagreement

We observe approval decisions y_{is} for each individual i in round s . The goal is to bound disagreement δ_{AB} , the share of claimants that first-round judges A_1 and B_1 would disagree on. Using 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

$$\delta_{AB} \equiv P[y_{i1}(A_1) \neq y_{i1}(B_1)] \quad (1)$$

It is a function of the unobserved joint distribution of $(y_{i1}(A_1), y_{i1}(B_1))$ and so is only partially identified. Defining D_s^j as a dummy variable indicating assignment to judge j , Fischman (2013) shows that

$$\begin{aligned} \delta_{AB} &= P[y_{i1}(A_1) = 1, y_{i1}(B_1) = 0] + P[y_{i1}(A_1) = 0, y_{i1}(B_1) = 1] \\ &\geq \max\{0, P[y_{i1}(A_1) = 1] - P[y_{i1}(B_1) = 1]\} + \max\{0, P[y_{i1}(B_1) = 1] - P[y_{i1}(A_1) = 1]\} \\ &= \max\{P[y_{i1}(A_1) = 1] - P[y_{i1}(B_1) = 1], P[y_{i1}(B_1) = 1] - P[y_{i1}(A_1) = 1]\} \\ &= \max\{E[y_1|D_1^A] - E[y_1|D_1^B], E[y_1|D_1^B] - E[y_1|D_1^A]\} \end{aligned} \quad (2)$$

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

The Equation 2 bound is the level of disagreement implied by the examiner-assignment monotonicity assumption. It is uninformative when judges A_1 and B_1 have the same approval rate. However, the same bound also applies within each demographic subgroup; if judges are in perfect agreement, they should approve the same share of cases within each demographic cell. Exploiting additional information about covariates can therefore make bounds informative even between judges with the same approval rate. For example, suppose that half of claimants are black, and half white.

If judge A_1 approved 75% of blacks and 25% of whites, and judge B_1 approved 25% of blacks and 75% of whites, they must disagree on at least 50% of all cases.

Bounds can also be tightened using latent claimant characteristics, such as whether a claimant *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. If judges A_1 and B_1 are perfectly consistent, then in expectation $P[y_{i2}(C_2) = 1|y_{i1}(A_1)] = P[y_{i2}(C_2) = 1|y_{i1}(B_1)]$. The following theorem shows how deviations from this ideal puts bounds on disagreement.

Theorem 1 (Bounding disagreement). *Suppose that $y_{i2}(C_2) \perp D_1^j$, 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:*

$$\begin{aligned} \delta_{AB}(C) &\geq \delta_{AB}^l(C) \\ &= \max \left\{ \begin{aligned} &E[y_1|D_1^A] - E[y_1|D_1^B], \\ &E[y_1|D_1^B] - E[y_1|D_1^A], \\ &E[y_1|D_1^A] - E[y_1|D_1^B] + 2 \left[E[y_1|D_1^B]E[y_2|D_1^B, D_2^C, y_1 = 1] - E[y_1|D_1^A]E[y_2|D_1^A, D_2^C, y_1 = 1] \right], \\ &E[y_1|D_1^B] - E[y_1|D_1^A] + 2 \left[E[y_1|D_1^A]E[y_2|D_1^A, D_2^C, y_1 = 1] - E[y_1|D_1^B]E[y_2|D_1^B, D_2^C, y_1 = 1] \right] \end{aligned} \right\} \end{aligned} \quad (3)$$

Proof: See Appendix Section A5.

Using Equation 3, we can do no worse than the Equation 2, highlighting the value of information about second-round decisions. Crucially, the last two terms in Equation 3 mean that the bound can be informative even when judges A_1 and B_1 have the same first round approval rate. This bound holds for all potential second-round judges, and can be combined with a max function. In other words, a weakly more informative bound on disagreement is

$$\delta_{AB} \geq \delta_{AB}^l = \max_j \{ \delta_{AB}^l(j) \} \quad (4)$$

I conduct estimation and inference of the disagreement bounds in Equation 4 using Chernozhukov et al. (2013). I report half-median unbiased estimates of δ_{AB}^l for each pair of judges, as well as the endpoints of one-sided 95% confidence intervals.⁷

⁷Half-median unbiased means that the estimate asymptotically falls below the true lower bound with probability at least $\frac{1}{2}$.

3 Structural model of judge decisions

In this section I introduce a structural model that explicitly connects inconsistency to a familiar index model of decision-making. The model has two main benefits. First, it clarifies the assumptions under which disagreement can be point-identified, rather than bounded. Second, for each judge it compresses their set of pairwise disagreements with each other judge into a two-parameter characterization of behavior, allowing examination of the determinants of judge quality.

The court receives applicants for refugee status, who are randomly assigned to judges. Strength of case for each applicant i can be represented by the scalar r_i , and represents the consensus decision opinion of all the judges. 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 $s \in \{1, 2\}$, judges $j \in \{1, \dots, J\}$ approve the claimant if

$$r_i > \varepsilon_{ijs}(X_{ijs}, W_{ijs}) = \gamma_{js} + X_{ijs}\beta_s + \tilde{\varepsilon}_{ijs}(W_{ijs}) \quad (5)$$

where

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 *standards* are captured by γ_{js} ; judges with higher values of γ_{js} approve fewer claimants. This threshold can be adjusted by covariates through the index $X_{ijs}\beta_s$, for example to allow changing standards as a function of experience.

Judge *accuracy* with respect to the court's collective standards is defined by the distribution of $\tilde{\varepsilon}_{js}(W_{ijs})$. For perfectly accurate 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, as in the standard index model of decision-making. Furthermore, both judges would approve the claimants most likely to be subsequently approved in the second round. As the distribution of $\tilde{\varepsilon}_{js}(W_{ijs})$ widens, judges deviate more from the collective standards.

3.1 Ranking judges

When can information about all $J(J-1)/2$ judge-pair disagreement rates be aggregated up into a ranking of all judges? Heuristically, better judges approve more high-quality claimants. To ensure that this concept is well-defined—for example, that there is not a judge A who approves more very high- r_i claimants than judge B, but also more very low- r_i claimants—requires restricting the distribution of judge errors. The following assumption applies to the set of *comparable* judges, who have the same first-round approval rates.

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:*

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

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

I define the more accurate judge as the judge approving more high- r_i claimants: judge A if condition (b)(i) is met, and judge B if (b)(ii).⁸

Note the contrast to the approach in [Section 2](#), where disagreement between pairs of judges does not have a normative interpretation. Here, the index representation of the judge decision in [Equation 5](#) and [Assumption 1](#) takes the normative stance that the judge who approves higher- r_i claimants—who are more likely to be approved in the second round—is the more accurate judge.⁹

It is simple to see that the ranking is transitive under the restricted set of distributions G_{js} that satisfy single-crossing for all comparable judge pairs; if judge A is more accurate than judge B, and judge B more accurate than judge C, then judge A is more accurate than judge C. This allows ranking all judges on a single scale. In the empirical application I implement this restriction by assuming that G_{js} is normally distributed.

⁸[Sah and Stiglitz \(1986\)](#) study a similar definition of judge quality, where they refer to the better judge as more discriminating.

⁹My model of decision-making has parallels with [Abaluck et al. \(2016\)](#)’s model of how doctors choose patients to send for further tests. Since the test reveals whether the patient actually has the disease, high test yield rates conditional on the share of patients sent for a test indicate a 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 accuracy.

3.2 Identification

The model—parameters β_s , judge-round thresholds γ_{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, and regressors that shift judge standards. I discuss each in turn.

3.2.1 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 accurate judge’s claimants will be ultimately approved by a common second-round judge, C_1 . Suppose that judge A_1 is more accurate. 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:

$$\begin{aligned}
& \left(P[r > \varepsilon_{C2} | r > \varepsilon_{A1}] - P[r > \varepsilon_{C2} | r > \varepsilon_{B1}] \right) P[r > \varepsilon_{B1}] \\
&= P[r > \varepsilon_{A1} \cap r > \varepsilon_{C2}] - P[r > \varepsilon_{B1} \cap r > \varepsilon_{C2}] \\
&= \int_{-\infty}^z [\tilde{G}_{A1}(r) - \tilde{G}_{B1}(r)] \tilde{G}_{C2}(r) f_r dr + \int_z^{\infty} [\tilde{G}_{A1}(r) - \tilde{G}_{B1}(r)] \tilde{G}_{C2}(r) f_r dr \\
&\geq \int_{-\infty}^z [\tilde{G}_{A1}(r) - \tilde{G}_{B1}(r)] \tilde{G}_{C2}(z) f_r dr + \int_z^{\infty} [\tilde{G}_{A1}(r) - \tilde{G}_{B1}(r)] \tilde{G}_{C2}(z) f_r dr \\
&= \tilde{G}_{C2}(z) \int [\tilde{G}_{A1}(r) - \tilde{G}_{B1}(r)] f_r dr \\
&= 0
\end{aligned} \tag{6}$$

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 the CDF $\tilde{G}_{C2}(\cdot)$ in r_i . Case strength r_i is the index of characteristics that improve likelihood of success in the first and second rounds, and so the judge identification tells us which first-round judge is approving higher-quality case in this sense.

Identification of second-round accuracy follows a similar route. Because I focus on first-round accuracy in the empirical results, I defer discussion of judge-assignment identification in the second round to [Appendix A6](#).

3.2.2 Regressors and identification

The between-judge comparisons in the previous section measure relative accuracy for judges with similar approval rates. To compare judges who approve different shares of claimants and to identify the scale of judge errors $\tilde{\varepsilon}_{ijs}$, additional large-support continuous regressors X_{ijs} are required. These regressors affect judge severity through standards γ_{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 $\tilde{\varepsilon}_{ijs} - r_i$ 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 $r_i, \tilde{\varepsilon}_{ij2}$.
- (c) There exist some columns of X_{ijs} , referred to as X_{ijs}^* , that are not in W_{ijs} .
- (d) At least one element of β_1 associated with X_{ijs}^* is equal to the same element in β_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. In the Appendix, I follow [Chen, Heckman, and Vytlacil \(2000\)](#) and show that these conditions are sufficient for identification.

Theorem 2 (Identification). *Suppose that [Assumption 1](#) and [Assumption 3](#) are satisfied. Then, the coefficients γ_{js} and β_s , and the distributions F_r and $G_{js,W}$ are identified.*

Proof: See Appendix Section A7.

3.3 Interpretation

In this section I introduce an additional condition that point-identifies disagreement. I then discuss the relationship between accuracy and substantive decision quality.

3.3.1 Point-identification of disagreement

Disagreement is a function of the joint distribution of errors for same-round judges. The joint distribution is not point-identified, so we need an additional assumption with an immediate implication.

Theorem 3 (Point-identifying disagreement). *Suppose that cross-judge errors within the first round are independent, or $\tilde{\varepsilon}_{ij1} \perp \tilde{\varepsilon}_{ik1} \forall j \neq k$. Then, disagreement is point-identified as $\delta_{jk} = \int [G_{j1}(r) + G_{k1}(r) - 2G_{j1}(r)G_{k1}(r)]f_r dr$.*

While the assumption of independence of first-round judge first-round errors is not directly testable—because different judges do *not* observe the exact same first-round cases—it is more plausible when the first and second rounds are similar, as in this context. This is because the model

guarantees that the first and second round errors are independent. Nonetheless, it would be useful to be able to test a more direct implication of the assumption in [Theorem 3](#), which I do in the following section.

3.3.2 Testing index sufficiency

Under the index model of decision-making, judges agree in principle on the ranking of cases by quality r_i , but differ in their ability to observe that quality. An alternative interpretation of the data is that different judges have different—but potentially legitimate—preferences that they observe perfectly. For example, half of the judges might care only whether the correct procedure had been followed during the initial decision, while the other half cares only about whether the claimant is a genuine refugee.

The distinction is important, because judges having preferences over different attributes of cases violates the within-round independence assumption that point-identifies disagreement ([Theorem 3](#)). Whether certain pairs of judges differentially prefer the same types of claimants (ie, $\tilde{\varepsilon}_{ij1}$ is correlated with $\tilde{\varepsilon}_{ik1}$) is not directly testable, but an indirect test is to ask whether there is a correlation in preferences by different judges *across* rounds (ie, whether $\tilde{\varepsilon}_{ij1}$ and $\tilde{\varepsilon}_{ik2}$ are correlated).

The test follows from the observation that second-round approval depends on the correlation between the first- and second-round judges errors. Defining

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

where ρ is the correlation between $\tilde{\varepsilon}_{ik1}$ and $\tilde{\varepsilon}_{ij2}$, $P_{jk}(\rho)$ is increasing in ρ .

By [Assumption 1](#), $\rho = 0$ for each judge pair. However, the parametric model that I estimate summarizes judge behavior as depending on only two parameters in each round, so there are more pairs of judges ($J(J-1)/2$, not counting the same judge in both rounds) than parameters determining $P_{jk}(\rho_{jk})$ ($4J$). To test the null hypothesis of $\rho = 0$, I conduct an overidentification test for residual judge-pair correlations, regressing

$$\mathbb{1}[\text{Approval by } j | \text{Approval by } k] = \beta \hat{P}_{jk}(0) + \nu_{jk} + u_{ijk} \quad (7)$$

where $\hat{P}_{jk}(0)$ is calculated using the estimated model parameters. Under the null, the judge-pair fixed effects ν_{jk} should be jointly insignificant. Failure to reject suggests that the assumption in [Theorem 3](#) of uncorrelated within-round judge errors is reasonable. I implement this test in [Section 4.7](#), and fail to reject the null of no judge-pair effects.

3.3.3 Model paramters, consistency and decision quality

Judge standards and accuracy are related to inconsistency in a simple way. Judges in a perfectly consistent court would all use the same threshold γ_{js} , and be perfectly accurate (e.g. $\tilde{\varepsilon}_{js} = 0$). Thus, any individual would be always approved by the court, or always rejected. Cross-judge variation in γ_{js} reduces consistency by increasing the variation in outcomes generated by judge assignment. Similarly, ordering errors induce randomness into the realized outcome for each individual; consistency is monotonically increasing in judge accuracy, holding approval rates fixed.

The relationship between judge accuracy and the quality of decisions in a more abstract sense depends the institution. Accuracy increases the likelihood of approval for high- r_i claimants at the expense of low- r_i claimants; the relationship between judge accuracy and decision quality 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. If the court is doing a good job on average—truly deserving claimants, however defined, are more likely to be approved than undeserving claimants—then higher levels of accuracy will also improve the quality of decisions, and accuracy is an appropriate measure of judge quality.

3.3.4 1st vs 2nd round judge errors

Interpreting errors in the second stage is complicated by the possibility that the judges gain additional information about the case in the second round (see [Section 1.2](#) for institutional details). I model this by making a distinction between idiosyncratic errors e_{ijs} and information \mathcal{I}_{ij2} , decomposing second-round errors into

$$\tilde{\varepsilon}_{ij2} = e_{ij2} + \mathcal{I}_{ij2} \quad (8)$$

The information shock \mathcal{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 \mathcal{I}_{ij2} .

The potential presence of \mathcal{I}_{ij2} complicates the interpretation of cross-judge differences in second-round ordering, because some of the differences may reflect new information, rather than an error. A wide distribution of $\tilde{\varepsilon}_{ij2}$ could reflect a particularly perceptive judge, so the relationship between judge quality and second-round errors could go in either direction. For this reason, I focus my discussion on first-round errors $\tilde{\varepsilon}_{ij1}$.

4 Data and results

My main data come from online Federal Court case reports.¹⁰ The case reports contain information on the date the case was filed, the office that received the application, the names of the assigned judges, and the outcome in each round. To collect information on the claimant gender and country of origin, I link the court records to the subset of available IRB case files.

For each judge, I collected information on the date and party of appointment. Appendix Table 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 period, 72% of judges are Liberal appointees. The average judge has 6.5 years of experience with a maximum of 28.

4.1 Estimation

Nonparametric identification of the structural model as outlined in Section 3.2 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 $\tilde{\varepsilon}_{ij1}$ and $\tilde{\varepsilon}_{ij2}$, generating tractable analytic expressions for approval probabilities (available in Appendix Section A8). I begin by assuming that $\tilde{\varepsilon}_{ijs}$ is mean-zero and normally distributed with standard deviation σ_{js} to be estimated as the measure of judge accuracy. Larger σ_{js} corresponds to less accuracy, ie a wider distribution of judge errors $\tilde{\varepsilon}_{ijs}$. I additionally allow regressors W_{ijs} to affect errors, so

$$\tilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2), \quad \sigma_{js}(W_{ijs}) = e^{\tilde{\sigma}_{js} + W_{ijs}\psi_s} \quad (9)$$

Since occasionally the same judge is assigned to the first- and second-round decision, I allow $\tilde{\varepsilon}_{ij1}$ and $\tilde{\varepsilon}_{ij2}$ to be correlated for all j (with the correlation estimated as an additional parameter).

As discussed in Section 1.2, the distribution of case quality is likely right-skewed, since it reflects claimants who were denied refugee status by government decision-makers. I therefore assume that r_i is distributed as an exponential-Pareto (I show the distribution of r_i relative to the estimated parameters in Figure 4). I allow flexibility in the distribution across two dimensions. First, cases filed at different offices likely vary in strength because different provinces vary in their level of legal aid funding. Second, the government made changes to the decision process for initial refugee applications in 2002 (see Section 1.1 for more details). Combining these, I fix the distribution of r_i to have location and shape parameters 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, but considerable between-office variation.

¹⁰ Available at http://cas-cdc-www02.cas-satj.gc.ca/IndexingQueries/infp_queries_e.php.

Nonparametric identification also requires a subset of regressors X_{ijs}^* that shift judge thresholds γ_{js} but do not affect judge errors $\tilde{\varepsilon}_{js}$. One immediate implication of this is that X_{ijs}^* is not in W_{ijs} , so that there is variation in X_{ijs}^* conditional on W_{ijs} .

For X_{ijs}^* I use the timing of the decisions during the week, and of the second-round hearing during the day, as regressors. [Danziger et al. \(2011\)](#) argue that when decision-makers make many decisions in a row, fatigue makes them more likely to pick the default option. They find that parole 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 second-round hearing as a regressor.

The second dimension of fatigue I exploit is over days of the week. Judges typically start working on first-round refugee cases at the start of the week, and work until they are finished. Because the day of decision is potentially endogenous—judges can choose the order they decide the cases—I use a dummy for the previous case action happening on a Wednesday or later, which predicts the first-round decision will happen on the second day of decisions or later. For second-round cases, which are scheduled by directly by staff, I directly use a dummy for the decision happening on the second or subsequent 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 A9.1 I show that these regressors are highly predictive of approval, suggesting that they lend identifying power in the spirit of the [Assumption 3](#) support conditions. However, they are uncorrelated with the likelihood of approval as predicted by fixed demographic characteristics of the claimants, suggesting they are orthogonal to r_i and $\tilde{\varepsilon}_{ijs}(W_{ijs})$. To further test this assumption, I estimate versions of the model where X_{ijs}^* , the excluded components of X_{ijs} , are added in turn to W_{ijs} —in other words, I allow the regressors to directly affect the level of judge accuracy. Since there are two regressors in the second round (dummies for the end of the week and a lunchtime hearing), I can test whether $\sigma_{j2}(W_{ij2})$ varies with the regressors and find that the effect is small and dominated by the effect through β_2 , in line with the assumption.

As a final robustness check, I estimate the model without regressors. This approach does not require [Assumption 3](#), but at the cost of relying on functional form for identification. In Appendix Section A9.3, I show that all of the results are qualitatively similar.

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

4.2 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, I regress claimant characteristics on judge-level mean approval rates, controlling for office \times pre-2002 fixed effects analogously to 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 significance.

The outcomes in Columns 1-5 are gender, region of origin, and the mean approval rate of the IRB Member that denied the claimants' initial application for refugee status. The predictive power of judge assignment is low, and the coefficients on the judge-level approval rate 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).

In Column 6, the outcome is an omnibus measure of claimant quality constructed by taking the predicted values from a regression of approval on the variables in the first five columns. Again, 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 also quasi-random.

4.3 Reduced form judge behavior

Variation in judge approval rates

Federal Court judges must show deference to the government's initial decision. 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%.

Evidence for common factor r_i

The dramatic increase in the approval rate from the first to the second round suggests that first-round judges successfully choose claimants who their colleagues agree have a strong case. In terms of the structural model, this implies that variation in refugee quality r_i is substantial. Judge preferences over claimants of different quality are also persistent; Panel C of Figure 1 shows that there is a strong relationship (correlation=0.56) judges' first- and second-round approval rates.

Another implication of a common factor r_i is that claimants approved by strict first-round 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 mean approval rate of 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 FEs. The results are similar; *having been approved* by a 10 pp more lenient first-round judge gives claimants a 2.6-3.2 pp lower chance of second-round approval.

Evidence on variation in judge accuracy

Accurate (low- σ_{js}) judges approve higher-quality claimants, conditional on approval rates. To generate reduced-form evidence on the level of variation in this ability, I calculate the mean approval rate in the second-round by the identity of the first-round approving judge, residualizing out the second-round judges approval rate and shrinking the estimates via Empirical Bayes to account for small cell sizes.

Figure 2 displays a scatter of this measure of judge ability against the judges first-round approval rate. As expected, the relationship is negative—stricter judges approve higher quality claimants, on average. However, there is a large degree of cross-judge dispersion in second-round approval for each first-round approval rate—the subsequent approval rate for judges approving about 15% of first-round applicants ranges from 38% to 48%. This implies a high degree of cross-judge variation in first-round accuracy.

4.4 Bounds on judge disagreement

Figure 3 displays the estimated bounds on disagreement for all pairs of first-round judges. I calculate these bounds in two ways: first, the naive examiner-assignment bound using only information on first-round approval rates (Equation 2), and second, using the additional information from second-round decisions (Equation 4).

Given that there are thousands of pairs, I show nonparametric regressions of the bounds and the corresponding confidence interval endpoints on the difference in first-round approval rates. As expected, the naive bound on disagreement is 0 for similarly-severe judges, but informatively different than zero under the full-information bound. In fact, the average disagreement rate for judge pairs with an approval rate within 1% is 8%, which is large given that the average approval rate is only 14%. Consistent with Figure 2, the judges make substantial ordering errors.

I also measure disagreement over all judge pairs. Using the additional information from second-round decisions, the average bound on disagreement can be tightened to 16.9%, with an average endpoint of the one-sided 5% CI of 11.9%.¹¹ This is substantially larger than the naive bound, which averages to 7.1%.

The degree to which judge disagreement is caused by ordering errors is captured by the difference between the naive and the full-information bound. To contextualize the degree of inconsistency, I also estimate the disagreement rate if judges decided cases by randomly approving claimants at their given approval rate. Under this benchmark of poor court performance, they would disagree on 25.1% of cases. In other words, the current court is at best halfway between the best- and worst-case scenarios.

¹¹I reject no disagreement for 73.8% of judge pairs at the 5% level without accounting for multiple testing, and 45.8% controlling for a false discovery proportion of 10% using the stepdown method of Romano and Shaikh (2006).

To explore the degree and causes of judge inconsistency further, for the rest of the paper I turn to the fully structural model of [Section 3](#). This model has the advantage of aggregating data up into a single, judge-specific measure of accuracy, as opposed to a judge-pair-specific measures of disagreement. That allows me to measure how accuracy and disagreement vary with judge covariates and other factors.

4.5 Structural results

The structural model estimates the parameters of the judges decision about whether to approve a claimant of quality r_i , displayed in [Equations 5](#) and [9](#) and reproduced here:

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

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

The main parameters of interest are γ_{j1} (judge-specific standards), σ_{j1} (judge-specific inaccuracy), and ψ_1 (the effect of covariates on judge accuracy). In the baseline model I concentrate on γ_{j1} and σ_{j1} , and estimate the model with an empty W_{ijs} .

Judge standards γ_{js} and inaccuracy σ_{js}

[Figure 4](#) plots the distribution of judge-round thresholds γ_{js} and inaccuracy σ_{js} . The red dotted 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 et al., 2008](#)).

In Panel A, the distribution of γ_{j1} is large relative to the distribution of refugee strength r_i , plotted in black. This means that not all variation in approval rates across judges comes from differences in accuracy—there are real differences in standards across judges. If judges were perfectly accurate ($\sigma_{j1} = 0$), the harshest judges would approve almost no claimants, and the most lenient almost all claimants.

Panel C shows the distribution of second-round standards γ_{j2} . They are slightly higher than the first-round standards, with an average of 2.99 versus 2.27 in the first round. This is consistent with optimal information-gathering, where first round judges approve some cases they are unsure about so their colleagues can take a second look with a higher standard.

Panel B of [Figure 4](#) shows the distribution of first-round inaccuracy σ_{j1} . The standard deviation of errors for the median first-round judge is 0.99, which is large relative to the standard deviation of 1 for underlying case quality r_i . As with the bounding approach, the ordering errors contribute meaningfully to the share of cases that judges disagree on. I calculate that the average pair of judges disagrees on 17.3% of cases, slightly larger than the 16.9% bound and large considering that

only 14% of cases are approved.

The distribution of σ_{j1} in Panel B is wide—some judges are much more accurate than others—so realistic improvements in accuracy dramatically reduce disagreement. If the least-accurate half of judges were improved to the level of the median judge, disagreement would fall by 15%, from 17.3% to 14.8%.

In Panel D of Figure 4, the distribution of second-round inaccuracy σ_2 is harder to interpret. Recall from Section 3.3.4 that the second-round error may contain an informational component \mathcal{I}_{ij2} . If the informational component is large, then a larger σ_{j2} may reflect better information-gathering abilities rather than fewer observational errors. Consistent with this interpretation, the correlation between σ_{j1} and σ_{j2} is -0.18. A negative correlation suggests that the low-error first-round judges are better at information-gathering, muddying the interpretation of second-round errors σ_{j2} .¹² I take this as a further reason to focus on first-round consistency for the remainder of the paper.

Quality of court decisions

Claimant quality r_i reflects the judges' aggregate opinion about which claimants have the most compelling cases. An important question is how often the current system approves those claimants.

In Figure 5 I provide evidence that first-round judges struggle to predict which claimants will be successful in the second round. For each r_i I calculate the first-round approval probability and the second-round approval probability conditional on first-round approval, and plot them against each other. The median claimant has a 5% chance of first-round approval, but a 19% probability of approval in the second round. This does not reflect selection, which is accounted for by conditioning on r_i . Instead, it shows that second-round decisions are unpredictable from the perspective of the first round, and even claimants with a relatively low r_i are sometimes approved in the second round.

Furthermore, no one has a very high chance of approval: the 95th percentile claimant has only a 67% chance of first-round approval and a 57% chance of second-round approval, for a 41% probability of overall success. 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 7.5% (0.14×0.52) from its current value of 6%.

In contrast, if all claimants were automatically approved in the first round, 24.2% of claimants would be successful in the second round. This reflects the relatively weak relationship between the likelihood of first- and second-round approval, and means that only one quarter of the claimants who would be successful in the second round are actually approved. The result is foreshadowed by Figure 2, which shows that the most lenient first-round judge approved 66.7% of claimants in the first round, and of those 27.4% were approved in the second round. Together, these approval shares bound the overall approval rate if there were no first round to at least 18.3% (0.667×0.274).

¹²Decomposing the first and second round residuals, we know that $\text{corr}(\text{var}(e_{ij1}), \text{var}(e_{ij2}) + \text{var}(\mathcal{I}_{ij2})) < 0$, so if $\text{var}(e_{ij1})$ and $\text{var}(e_{ij2})$ are positively correlated then $\text{corr}(\text{var}(e_{ij1}), \text{var}(\mathcal{I}_{ij2}))$ must be negative.

Although judges might change their behavior if the first round was abolished, this shows that the court is currently approving only a small fraction of the claimants who meet its standards.

4.6 Relationship between structural parameters and reduced form moments

In this section, I show how reduced-form moments translate into the first-round structural parameters (for second-round parameters, see Appendix Section A9.2). The principal determinant of first-round approval rates is judge thresholds γ_{j1} . In Panel A of [Figure 6](#), I vary the γ_{j1} 's from their estimated values by adding a common shifter to each γ_{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 3.2.1](#), the frequency with which approved claimants are subsequently approved in the second round helps identify first-round inaccuracy σ_{j1} . In Panel B, I adjust the σ_{j1} 's away from their estimated values by multiplying each coefficient by a common factor. As accuracy decreases (σ_{j1} gets larger), this dramatically reduces the second-round approval rate.

Panel C of [Figure 6](#) demonstrates the judge-comparison identification more directly. Comparing two judges with the same first-round approval rate, [Equation 6](#) indicates that the more accurate 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 pp of each other, then plot the difference in second-round approval rates against the difference in estimated σ_{j1} . As expected, judges with a higher second-round approval rate than their matched colleague have a lower estimated σ_{j1} , or are more accurate.

4.7 Testing uncorrelated judge errors

The key assumption to point-identify disagreement is that the scalar quality factor r_i is a sufficient measure of case strength, and conditional on r_i judges' errors are uncorrelated ([Theorem 3](#)). In [Section 3.3.2](#), I discuss how one implication of this assumption is that 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 ν_{jk} , the fixed effects should not 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.

Across columns, the coefficient on $P[\text{Approval by } j_2 | \text{Approval by } k_1]$ is statistically indistinguishable from one, but the judge-pair fixed effects are jointly insignificant with a p -value of about 0.7.¹³ As a descriptive analysis, I calculate the Empirical Bayes judge-pair means of the regression residuals. This confirms the F-stat result: the standard deviation of the judge-pair means is in the

¹³Since the F-test can over-reject when the judge-pair cells are small, I follow [Abrams et al. \(2012\)](#) and bootstrap the distribution of the null. The asymptotic p -values range from 0.30 to 0.60.

range of 0.001-0.006 relative to a mean approval rate of 0.44. Both pieces of evidence suggest that there is in fact no correlation between errors $\tilde{\varepsilon}_{ijs}$ for cross-round judge pairs.

The judge-pair test 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 7](#) with interactions of judge demographic characteristics, and test whether observable groups of judges share preferences for certain types of claimants. I conduct this test in [Table A2](#) for judge experience (more than 1 and more than 5 years of experience), gender, political party of appointment (Liberal or Conservative), and native language (French or English). 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 0.15 to 0.88. Taken together, these two tests imply that the assumption in [Theorem 3](#) is reasonable.

4.8 Judicial accuracy by experience and workload

The key input for optimizing judicial terms is the rate of learning. If judges learn slowly, that suggests that judicial churn is costly and should be avoided. [Table 4](#) presents models where I allow experience to directly affect judge accuracy $\sigma_{js}(W_{ijs})$ as a component of W_{ijs} . I parameterize experience with indicators for more than 1, 5, and 10 years of experience.

Column 1 shows that first-round inaccuracy σ_{j1} shrinks dramatically after the first year (0.77 log points), followed by smaller but still substantial decreases of 0.39 log points after five years and 0.41 log points after ten years. To put these numbers into context, if all judges had less than one year of experience, pairs of judges would disagree on 21.3% of cases, close to the random-selection benchmark of 25.1%. After one year of experience that declines to 16.7%; after 5 years, 15.3%; and after 10 years 13.1%.¹⁴

Front-loaded gains to experience are often observed in other contexts. However, in contrast to teachers—who see most of their gains after the first year—judges continue to improve for at least 10 years, perhaps reflecting the more complicated nature of the job ([Rivkin et al., 2005](#)).

One likely form of learning is younger judges slowly adjusting to the standards set by their older colleagues. I see some evidence of this in [Table A2](#), where I regress actual second-round approval on the model-predicted probability of approval as well as interactions between the first- and second-round judges level of experience (see [Section 4.7](#) for details). Above and beyond what is predicted by the model, first-round judges with less than one year of experience are 9.2 pp *less* likely to have their approvals subsequently approved by experienced second-round judges. In contrast, for experienced first-round judges, subsequent approval does not depend on second-round judge experience. This is consistent with experienced and inexperienced judges operating on slightly different standards, and

¹⁴In each counterfactual I adjust γ_{j1} to keep judge approval rates the same.

new judges learning those standards over their first year on the job.

Gains from assigning more cases to experienced judges

One implication of high returns to experience is that more cases should be given to experienced judges. There are two potential issues with this policy: first, that it might decrease the rate of learning for new judges; and second, that the extra workload might make the experienced judges less accurate. I address each of those issues in turn.

There are prior reasons to believe that judge accuracy will improve as a function of years of experience, not number of cases. First-round decisions are not written up, and the second-round decision happens much later (the median wait is 89 days), making it difficult to directly learn which cases are likely to be subsequently approved. Most learning about how their colleagues make first-round decisions is indirect: inferring them from the cases they observe themselves in the second round, and discussing cases with colleagues.

I test the form of learning in Column 2 of [Table 4](#), allowing years of experience *and* career number of cases to affect accuracy. After controlling for log number of cases, the returns to experience are actually slightly higher, with reductions of 1.15, 0.52, and 0.75 log points after 1, 5, and 10 years of experience. This suggests that judges improve over time rather than with the number of cases, and so assigning more cases to experienced judges would not harm the skill development of new judges.

The second potential issue with assigning more cases to experienced judges is that their productivity might drop, negating the gains. To test this, I calculate monthly log workload as the number of first-round cases a judge is assigned in a given month and add it to the vector of variables that affect judge thresholds (X_{ijs}) and errors (W_{ijs}). Columns 3 of [Table 4](#) shows that this indeed reduces accuracy; the elasticity of σ_{j1} with respect to workload is 0.15. However, this cost comes entirely from younger judges: in Column 4 I interact the workload variables with experience, and find that the workload elasticity is 0.31 for judges with less than 5 years of experience, but 0.03 for experienced judges. This means that transferring the marginal case from an experienced to an inexperienced judge improves average accuracy. Larger changes have the potential to moderately improve consistency; transferring one third of the inexperienced judges' cases to the experienced judges would reduce average disagreement by 0.8 pp.

4.9 Judge selection reform and judge consistency

Starting in 1988, a new reform required that candidates for judgeships be approved by an independent committee (see [Section 1.3](#) for details). The reform successfully reduced the number of new judges with ties to the ruling party. In this section, I show that it also improved judge accuracy.

[Table 5](#) presents regressions of $\hat{\sigma}_{j1}$ on a dummy for whether the judge was appointed before the reform, using both the baseline $\hat{\sigma}_{j1}$ estimate, and the [Table 4](#)-Column 1 version that residualizes out

experience.¹⁵ Accounting for experience turns out to be important, since the reform took place seven years before the start of my sample, making the pre-reform judges mechanically more experienced.¹⁶ I weight the regressions to account for estimation error in the dependent variable (Hanushek, 1974), and in my preferred specifications control for a year trend, judge gender and party of appointment

The first three columns of Table 5 show that average inaccuracy declined by 0.66 after the reform, a 65% reduction (p -value=0.11). In Columns 4-6, the effect is much more precise once the model properly accounts for experience, with the reform reducing σ_{j1} for new judges by a statistically significant 1.19 points relative to a pre-reform mean of 1.72.¹⁷

The substantial improvements in accuracy led to large improvements in consistency. Using the preferred right-most model, I calculate that the accuracy improvements from the reform dramatically decreased disagreement, from an average of 20.6% to 13.2%. This reflects the strength of the reform, which was stringent enough to restrict the Minister’s options to only 40% of the original set of applicants for judge positions.

Interestingly, the reform had no effect on judge leniency. In Appendix Table A3, I show that judges appointed after the reform approved a statistically indistinguishable share of first-round claimants. More directly, in Appendix Table A4 I control for judge approval rates and find similar results, with an average σ_{j1} 1.14 smaller for judges appointed after the reform (SE=0.279).

The table also shows that there are no substantial differences in accuracy by judge gender or party of appointment. The Federal Court is a prestigious appointment, and so neither Liberal or Conservative governments are likely to be constrained by supply limitations.

4.10 Judge errors and expert opinion

In this section, I study the relationship between the judge patterns and lawyer perceptions of judge ability. Correlation between model-based measures and lawyers’ opinion serve as a validation of the model, and suggests that other participants in the legal system have useful information about the quality of decisions. In the following section I will show that this information could be used by court administrators to optimally assign judges to cases.

To measure expert opinion, I conducted an email survey of refugee lawyers who have appeared at the Federal Court. I asked respondents to rate judges with whom they had personal experience along dimensions analogous to the parameters of the model: first, leniency (corresponding to judge

¹⁵For approximate comparability, I adjust all coefficients to the median experience of 6 years.

¹⁶One potential issue with not observing data at the time of the reform is that accurate judges might be more likely to be promoted to a higher court, making pre-reform judges look less accurate. Although about 20% of new judges are promoted in seven or fewer years, judges who are eventually promoted have an estimated σ_{j1} only 0.013 higher (standard error 0.13) than non-promoted judges, relative to a mean of 0.94.

¹⁷One alternative explanation for higher accuracy among post-reform judges would be a particular violation of index sufficiency, where pre- and post-reform judges value a different set of attributes. I test this using the characteristics test in Table A2, and find no evidence that pre- and post-reform judges are operating on different standards.

threshold γ_{js}); and second, consistency and predictability conditional on approval rate (I reverse this scale so more unpredictability corresponds to higher σ_{js}). More details about the survey are in Appendix Section A10.

Each response is on a five-point likert scale, which I normalize by the mean and standard deviation. Table 6 describes the relationship between model coefficients and the survey results. I model the relationship as

$$\hat{C}_j = \beta_0 + \beta_1 \text{Favorability}_{j\ell} + \beta_2 \text{Unpredictability}_{j\ell} + \eta_\ell + u_{j\ell}$$

where ℓ indexes lawyers and $\hat{C}_j = \{\hat{\gamma}_1, \hat{\sigma}_1\}$. I use model estimates that account for experience (which is highly predictive of behavior), and for each judge-respondent pair use the predicted parameter at the time of their modal interaction. To account for estimation error in the model coefficients I use Hanushek’s (1974) efficient weights.

In Columns 1 to 3, 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 right-most preferred specification a SD increase in favorability decreases γ_1 by 0.31, which is just shy of significance at the 10% level and about 0.38 SD of the cross-judge distribution of γ_1 .

Columns 4 to 6 show the relationship between surveyed judge characteristics and $\hat{\sigma}_1$. Survey unpredictability predicts lower model accuracy; across specifications one extra SD of surveyed unpredictability translates to 0.11-0.12 higher $\hat{\sigma}_1$, or about 0.27 SD of the cross-judge distribution. In other words, the model and the lawyer survey describe the same judges as being accurate, suggesting that the survey is picking up true variation in judge ability to assess case strength.

4.11 Optimal judge allocation

The goal of the court is to approve the strongest cases, implicitly subject to a budget constraint. With model estimates of judge accuracy, both the cost and quality dimensions of the goal can be made explicit. In this section I study the problem of minimizing court costs while approving the same number of claimants with strong cases.

Court costs are driven by the number of second-round hearings, which are much more time-intensive than first-round decisions. Instead of reading documents from the government’s initial decision, second-round decisions require an in-person hearing and a full written decision, which together take about ten times as long as a first-round decision. Thus, reducing the number of first-round approvals (and resulting second-round decisions) is the key to reducing the Court’s workload. The challenge is to do so without reducing the quality of the ultimately-approved claimants.

To operationalize this requirement, I minimize total workload over the set of assignments to pairs of first- and second-round judges that result in a posterior distribution of r_i for approved

claimants that first-order stochastically dominates the baseline distribution, and that approves at least as many cases. I also require that no judge works more than she currently does.

I conduct the optimization, and find that the optimal assignment rule would reduce total workload by 18.3%, amounting to savings of approximately \$4.6 million in judge salaries alone over the study period while allowing claimants to receive their decision faster.

Under this problem, there are three potential ways to minimize workload: (1) reallocating judges to rounds where they make more accurate decisions; (2) moving strict judges to the first round to reduce the number of second-round decisions; and (3) moving lenient judges to the second round so that the overall approval rate remains the same. The optimal solution shows the second two kinds of savings, but not the first. [Figure 7](#) presents histograms of the optimally-assigned judge coefficients, as well as histograms of the baseline allocation. The average first-round threshold γ_{j1} for optimally-assigned judges is higher (Panel A), meaning that there will be fewer, higher-quality first-round cases approved. Conversely, second-round judges have much lower standards γ_{j2} (Panel C). In Panel B and D, the change in the distribution of accuracy is much less dramatic. Average σ_{j1} is almost unchanged, though there is much less usage of the least accurate second-round judges. On net, [Figure 7](#) shows that judge standards γ_{js} are the most powerful lever to affect case quality, but knowledge of judge accuracy is required to discipline the allocation process.

Optimizing judge assignment with second-round information

The cost minimization problem above uses r_i as the relevant measure of quality. Implicitly, this assumes that there is no informational component in second-round errors $\tilde{\epsilon}_{ij2}$. Another possible goal is to assume that the second-round judge receives accurate information in the second round, and minimize cost subject to posterior $r_i + \mathcal{I}_{i2}$ first-order stochastically dominating the baseline distribution. Under this specification, the optimal allocation is highly correlated with the first optimal allocation (0.55), though workload is reduced by 12.1% rather than 18.3%. It also satisfies the original constraints, so a cautious planner could implement this design and enjoy most of the gains of judge reallocation.

4.12 Implications for monotonicity in examiner-assignment designs

Judge disagreement over the ordering of claimants by quality has important implications for examiner-assignment research designs. This identification strategy uses random assignment of decision-makers to cases to generate random variation in treatment, then studies the effect of a treatment—such as incarceration, patent receipt, or being placed in foster care—on outcomes ([Bhuller et al., 2016](#); [Gaulé, 2015](#); [Doyle, 2008](#)).

The strongest form of monotonicity requires that all individuals are weakly more likely to be approved by each high-approval judge than a lower-approval judge. Differential ordering of cases by

quality for pairs of judges with similar leniency violates monotonicity, but the degree of inaccuracy does not directly indicate the extent of these violations. In Appendix Section A1, I introduce new tools related to the techniques in this paper that can be used to directly test monotonicity.

I distinguish between average monotonicity—which is required for linear IV to return a convex combination of treatment effects—and the monotonicity conditions required for estimating MTEs. I find that MTE-monotonicity is violated, and in simulations show that the bias can be large. In contrast, I do not reject average monotonicity.

5 Conclusion

Predictable and consistent decision-making institutions are important economic assets. Courts secure property rights, enabling long-term planning and growth (Craswell and Calfee, 1986; Porta et al., 1998). SSDI has many low-probability applicants whose labor market skills decay while waiting for a decision; improving SSDI consistency would likely dissuade some of those applicants from applying in the first place (Autor et al., 2015).

A basic measure of consistency is the frequency with which particular cases would be decided differently. A recent literature studying examiners in large institutions has documented many non-relevant characteristics of the environment—including the outdoor temperature, the outcome of a football game, and the judge’s decision in the previous case—that affect classification outcomes such as incarceration rates, implying that at least some cases’ outcomes depend on external factors.

In this paper I focus on a much larger contributor to inconsistency, cross-examiner disagreements on decisions. I introduce new tools to measure inconsistency arising from examiner disagreements, and study how it changes with experience, workload, and the method of examiner selection.

The fundamental challenge in measuring examiner disagreement is that we never see two examiners make a decision on the same case, and so never know whether they disagree on any particular case. Some progress has been made using bounds: if one examiner approves 40% of cases while his colleague approves 50%, they must disagree on at least 10% of cases (Fischman, 2013). The examiner-assignment monotonicity assumption is equivalent to assuming this lower bound is the true level of disagreement, and even under this extreme assumption there is evidence of moderate levels of disagreement.

I show that the examiner assignment bound can be considerably tightened with data from a multiple-stage decision process, where the decision only occurs with the agreement of two consecutive examiners. Whenever there is quasi-random assignment of the examiners in both rounds, the second-round decision can be used as a stratifying covariate: if two first-round examiners would approve the exact same cases, then the same share of their approvals should be subsequently approved by each second round examiner. Deviations from this ideal imply disagreement, with the bounds getting

tighter as more second-round examiners become available.

I implement the method using data on judicial review of initially-denied refugee claims at the Federal Court of Canada. Although the justices are experts in refugee cases, I uncover high levels of inconsistency. I bound disagreement for the average pair of judges to at least 17%, which is strikingly high given that they only approve 14% of cases overall.

The high levels of disagreement in this court do not come primarily from differences across judges in their overall approval rates—in fact, the examiner-assignment IV bound constructed using only approval rates is just 7%. Instead, most court inconsistency is caused by judges disagreeing in their ordering of the cases by quality. I build a structural model to understand the causes of this ordering disagreement, where I use the two-stage court structure to separately estimate judge standards—the lowest quality case they would accept—and judge accuracy, or the degree to which their ranking of cases differs from the consensus.

Cross-judge variation in accuracy is substantial. Replacing the least accurate half of judges with median-accuracy judges would reduce the disagreement rate by 15%, from 17.3% to 14.8%. However, accuracy improves dramatically after the first year, and continues to improve for at least the first ten years of experience, meaning that shifting cases from inexperienced to experienced judges reduces disagreement. This policy would improve accuracy even on the margin: decreasing workload improves accuracy for inexperienced judges, but the corresponding increase in workload has no effect on experienced judges accuracy.

Most dramatically, the method of judicial selection matters. A reform in the late 1980s reduced the government’s discretion in making appointments by requiring that all candidates be approved by an independent committee of legal experts. This had the intended effect of reducing the number of party supporters appointed as judges, but also dramatically improved judicial accuracy. Over time, I estimate the reform decreased disagreement from 21% to 13% while leaving overall approval rates unchanged.

Another way to improve court functioning is to reallocate judges to the rounds where they are most accurate. These potential gains are large; the court could reduce judge labor costs by 18% while approving applicants of higher quality as measured by the cross-judge consensus.

In the Appendix, I show how the methods in this paper can be adopted into direct tests of the monotonicity assumption in examiner-assignment designs. I decisively reject the most stringent form of monotonicity; at least in this context, the MTE identification assumptions are violated.

Finally, my research has direct implications for the Federal Court. Under current policy, 14% of all claimants proceed to the second stage and 6% of the total are eventually successful. However, first-round judges reject many claimants who would be successful in the second stage—if first-stage approval became automatic, overall success would increase to 24.2%. Over the 17 years from 1995, that difference amounts to approximately 10,400 families.

References

- ABALUCK, J., L. AGHA, C. KABRHEL, A. RAJA, A. VENKATESH, ET AL. (2016): “The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care,” *American Economic Review*, 106, 3730–3764.
- ABRAMS, D. S., M. BERTRAND, AND S. MULLAINATHAN (2012): “Do Judges Vary in Their Treatment of Race?” *The Journal of Legal Studies*, 41, 347–383.
- ANGRIST, J. D., G. W. IMBENS, AND D. B. RUBIN (1996): “Identification of causal effects using instrumental variables,” *Journal of the American statistical Association*, 91, 444–455.
- ARNOLD, D., W. DOBBIE, AND C. S. YANG (2017): “Racial bias in bail decisions,” *The Quarterly Journal of Economics*.
- AUTOR, D., N. MAESTAS, K. J. MULLEN, AND A. STRAND (2015): “Does delay cause decay? The effect of administrative decision time on the labor force participation and earnings of disability applicants,” Tech. rep., National Bureau of Economic Research.
- BERNANKE, B. S. (1983): “Irreversibility, uncertainty, and cyclical investment,” *The Quarterly Journal of Economics*, 98, 85–106.
- BHULLER, M., G. B. DAHL, K. V. LØKEN, AND M. MOGSTAD (2016): “Incarceration, recidivism and employment,” Tech. rep., National Bureau of Economic Research.
- CARD, D., A. MAS, E. MORETTI, AND E. SAEZ (2012): “Inequality at work: The effect of peer salaries on job satisfaction,” *The American Economic Review*, 102, 2981–3003.
- CHEN, D. L., T. J. MOSKOWITZ, AND K. SHUE (2016): “Decision Making Under the Gambler’s Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires,” *The Quarterly Journal of Economics*, 131, 1181–1242.
- CHEN, D. L. AND A. PHILIPPE (2018): “Clash of norms: Judicial leniency on defendant birthdays,” Tech. rep., Institute for Advanced Study in Toulouse (IAST).
- CHEN, X., J. J. HECKMAN, AND E. VYTLACIL (2000): “Identification and SQRT N Efficient Estimation of Semiparametric Panel Data Models with Binary Dependent Variables and a Latent Factor,” in *Econometric Society World Congress 2000 Contributed Papers*, Econometric Society, 1567.
- CHERNOZHUKOV, V., S. LEE, AND A. M. ROSEN (2013): “Intersection bounds: estimation and inference,” *Econometrica*, 81, 667–737.

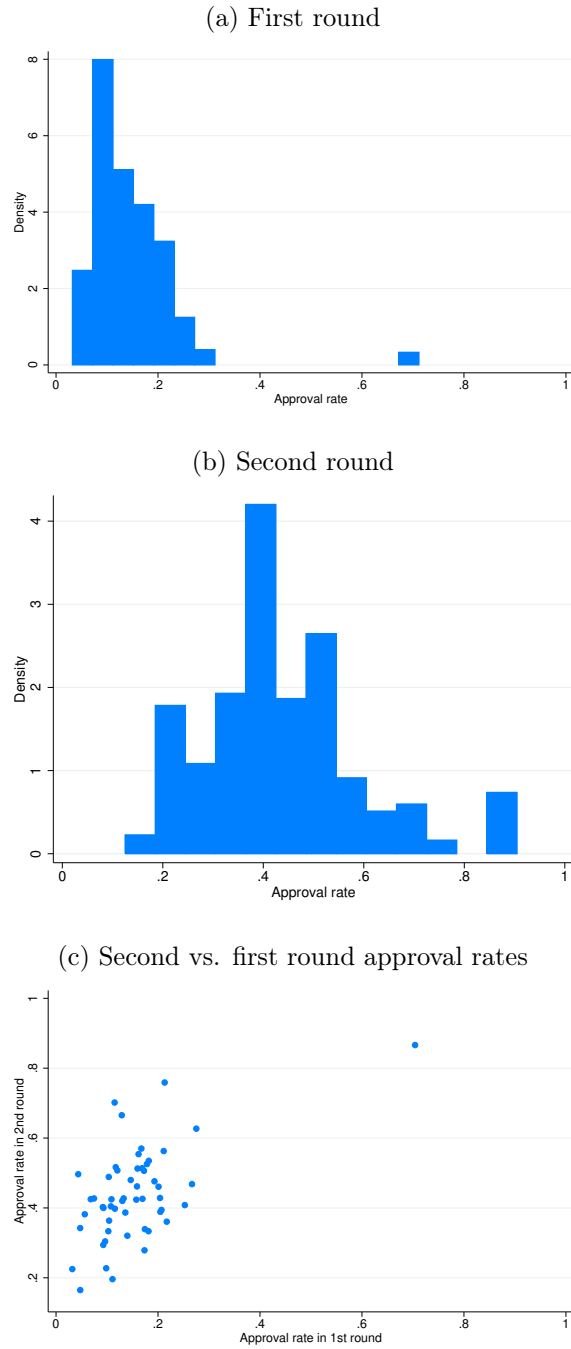
- CRASWELL, R. AND J. E. CALFEE (1986): “Deterrence and uncertain legal standards,” *JL Econ. & Org.*, 2, 279.
- DANZIGER, S., J. LEVAV, AND L. AVNAIM-PESSE (2011): “Extraneous factors in judicial decisions,” *Proceedings of the National Academy of Sciences*, 108, 6889–6892.
- DE CHAISEMARTIN, C. (2017): “Tolerating defiance? Local average treatment effects without monotonicity,” *Quantitative Economics*, 8, 367–396.
- DELAIGLE, A., P. HALL, AND A. MEISTER (2008): “On deconvolution with repeated measurements,” *The Annals of Statistics*, 665–685.
- DJANKOV, S., R. LA PORTA, F. LOPEZ-DE SILANES, AND A. SHLEIFER (2003): “Courts,” *The Quarterly Journal of Economics*, 118, 453–517.
- DOBBIE, W., J. GOLDIN, AND C. S. YANG (2018): “The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges,” *American Economic Review*, 108, 201–40.
- DOYLE, J. J. (2008): “Child protection and adult crime: Using investigator assignment to estimate causal effects of foster care,” *Journal of political Economy*, 116, 746–770.
- EREN, O. AND N. MOCAN (2016): “Emotional judges and unlucky juveniles,” Tech. rep., National Bureau of Economic Research.
- FISCHMAN, J. B. (2013): “Measuring Inconsistency, Indeterminacy, and Error in Adjudication,” *American Law and Economics Review*, 16, 40–85.
- FRANDSEN, B. R., L. J. LEFGREN, AND E. C. LESLIE (2019): “Judging Judge Fixed Effects,” Tech. rep., National Bureau of Economic Research.
- FRÉCHET, M. (1951): “Sur les tableaux de corrélation dont les marges sont données,” *Ann. Univ. Lyon, 3^e e serie, Sciences, Sect. A*, 14, 53–77.
- GAULÉ, P. (2015): “Patents and the success of venture-capital backed startups: Using examiner assignment to estimate causal effects,” .
- GRANT, A. G. AND S. REHAAG (2015): “Unappealing: An Assessment of the Limits on Appeal Rights in Canada’s New Refugee Determination System,” .
- HANUSHEK, E. A. (1974): “Efficient estimators for regressing regression coefficients,” *The American Statistician*, 28, 66–67.

- HAUSEGGER, L., T. RIDDELL, M. HENNIGAR, AND E. RICHEZ (2010): “Exploring the Links between Party and Appointment: Canadian Federal Judicial Appointments from 1989 to 2003,” *Canadian Journal of Political Science/Revue canadienne de science politique*, 43, 633–659.
- HECKMAN, J. J., S. URZUA, AND E. VYTLACIL (2006): “Estimation of treatment effects under essential heterogeneity,” *Health affairs (Project Hope)*, 29, 389–432.
- HECKMAN, J. J. AND E. VYTLACIL (2005): “Structural equations, treatment effects, and econometric policy evaluation,” *Econometrica*, 73, 669–738.
- HEYES, A. AND S. SABERIAN (2019): “Temperature and decisions: evidence from 207,000 court cases,” *American Economic Journal: Applied Economics*, 11, 238–65.
- KLEIN, T. J. (2010): “Heterogeneous treatment effects: Instrumental variables without monotonicity?” *Journal of Econometrics*, 155, 99–116.
- KLEINBERG, J., H. LAKKARAJU, J. LESKOVEC, J. LUDWIG, AND S. MULLAINATHAN (2017): “Human decisions and machine predictions,” *The quarterly journal of economics*, 133, 237–293.
- KLING, J. R. (2006): “Incarceration length, employment, and earnings,” *The American economic review*, 96, 863–876.
- MANSKI, C. F. (1975): “Maximum score estimation of the stochastic utility model of choice,” *Journal of econometrics*, 3, 205–228.
- MCKELVEY, S. (1985): “The Appointment of Judges in Canada,” Tech. rep., Canadian Bar Association.
- MUELLER-SMITH, M. (2015): “The criminal and labor market impacts of incarceration,” *Unpublished Working Paper*.
- NORRIS, S., M. PECENCO, AND J. WEAVER (2018): “The Intergenerational and Sibling Effects of Incarceration: Evidence from Ohio,” .
- PEW (2012): “Assessing the Representativeness of Public Opinion Surveys,” Tech. rep., Pew Research Center.
- PHILIPPE, A. AND A. OUSS (2016): ““No Hatred or Malice, Fear or Affection”: Media and Sentencing,” .
- PORTA, R. L., F. LOPEZ-DE SILANES, A. SHLEIFER, AND R. W. VISHNY (1998): “Law and finance,” *Journal of political economy*, 106, 1113–1155.

- RAO, C. R. (1971): “Characterization of probability laws by linear functions,” *Sankhyā: The Indian Journal of Statistics, Series A*, 265–270.
- REHAAG, S. (2007): “Troubling patterns in Canadian refugee adjudication,” *Ottawa L. Rev.*, 39, 335.
- (2012): “Judicial Review of Refugee Determinations: The Luck of the Draw?” .
- RIVKIN, S. G., E. A. HANUSHEK, AND J. F. KAIN (2005): “Teachers, schools, and academic achievement,” *Econometrica*, 73, 417–458.
- ROMANO, J. P. AND A. M. SHAIKH (2006): “On stepdown control of the false discovery proportion,” in *Optimality*, Institute of Mathematical Statistics, 33–50.
- RUSSELL, P. H. AND J. S. ZIEGEL (1991): “Federal Judicial Appointments: An Appraisal of the First Mulroney Government’s Appointments and the New Judicial Advisory Committees,” *The University of Toronto Law Journal*, 41, 4–37.
- SAH, R. K. AND J. E. STIGLITZ (1986): “The architecture of economic systems: Hierarchies and polyarchies,” *The American Economic Review*, 716–727.
- UNITED NATIONS (1967): “Protocol relating to the status of refugees,” *Treaty Series*, 30.
- WOLAK, F. A. (1989): “Testing inequality constraints in linear econometric models,” *Journal of econometrics*, 41, 205–235.

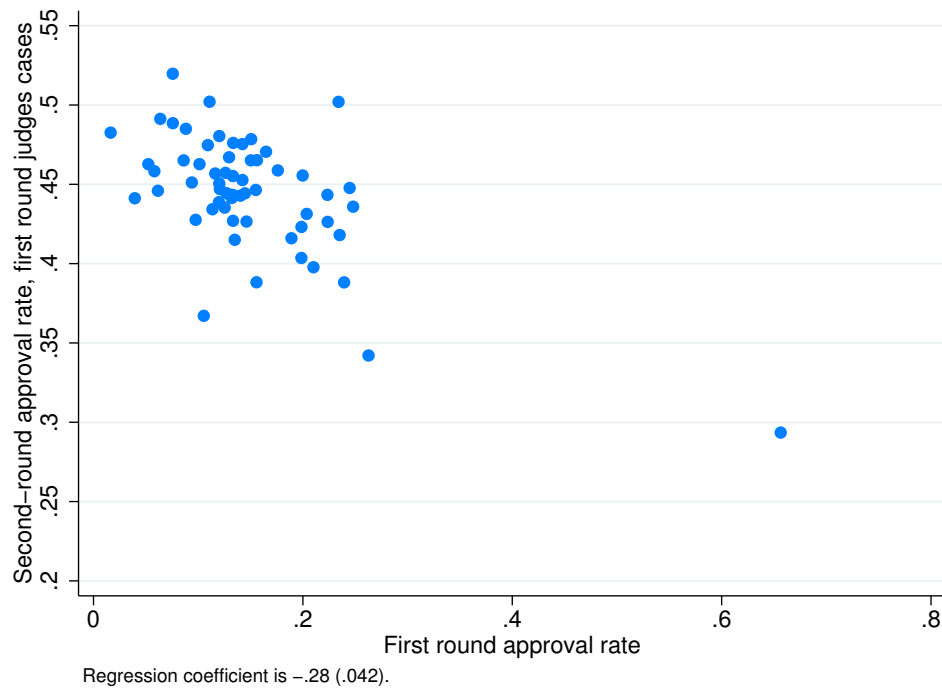
6 Figures

Figure 1: Approval rates by judge



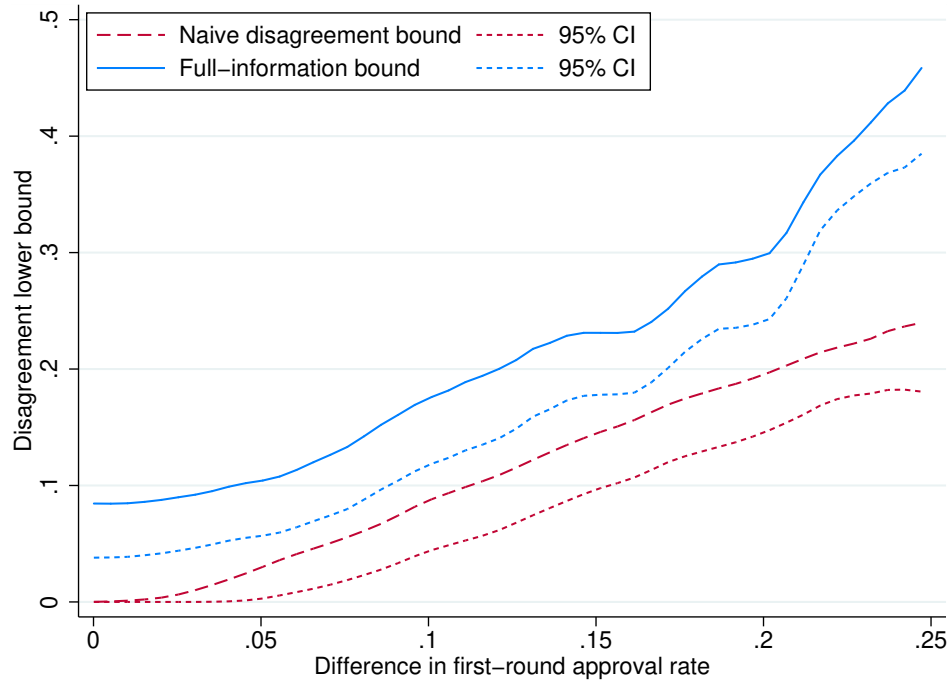
Panel A and B contain histograms of approval rates by judge for the first and second round, respectively. Both are weighted by the number of observations per judge. Panel C contains the scatter plot of judge-level first- and second-round approval rates. The correlation is 0.57, and 0.40 without the outlier. See [Section 4.3](#) for full discussion.

Figure 2: Share of approved claimants subsequently approved versus first-round approval rates



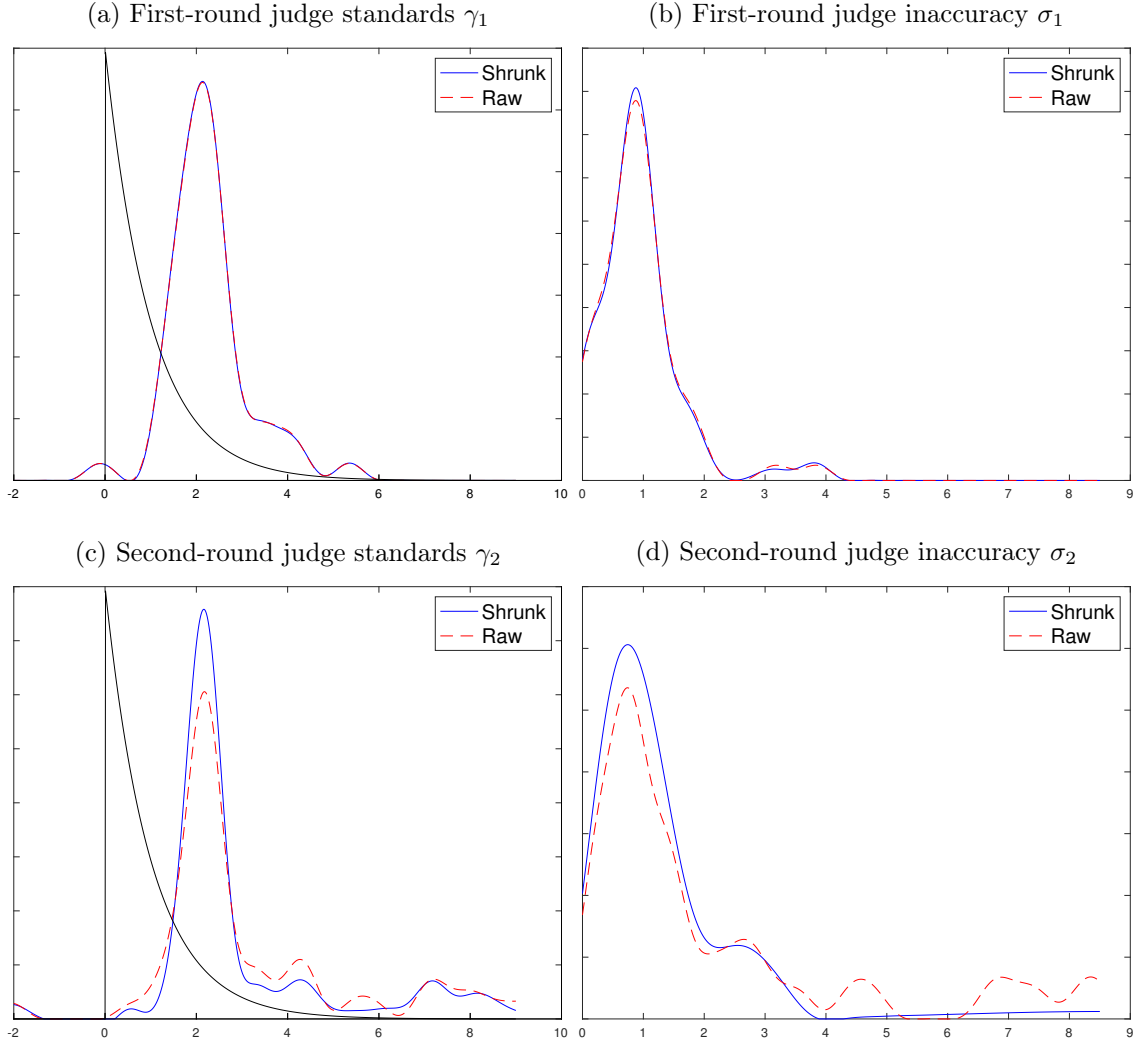
The figure shows second round approval rates for the claimants approved by each first round judge plotted against the judge's first-round approval rates, with second-round judge approval rates residualized out and means shrunk via Empirical Bayes to account for measurement error from small cells. See [Section 4.3](#) for full discussion.

Figure 3: Lower bounds for judge-pair disagreement rates by difference in first-round approval rate



This figure displays a local polynomial regression of the half-median-unbiased estimates of disagreement rates for pairs of first-round judges on the difference in first-round approval rates. The naive estimates use only variation in judge approval rates; the full-information estimates tighten the bounds using the outcome of the second-round decisions (Theorem 1). Dashed lines represent the lower endpoints of 95% one-sided confidence intervals. Estimates and CIs calculated using the method of Chernozhukov, Lee and Rosen (2013). See [Section 4.4](#) for full discussion.

Figure 4: Distribution of judge parameters



This figure presents coefficient estimates for the decision model $\mathbb{1}[r_i > \gamma_{js} + X_{ijs}\beta_s + \tilde{\varepsilon}_{ijs}]$, $\tilde{\varepsilon}_{ijs} \sim \mathcal{N}(0, \sigma_{js}^2)$. 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 shrunk estimates of the judge-round specific standards γ_1 and γ_2 , and inaccuracy σ_1 and σ_2 . Black line in Panels A and C is density of case quality r_i in base time period and office. Shrunk 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). See [Section 4.5](#) for full discussion.

Figure 5: Model estimates of first- vs. second-round approval, by case strength

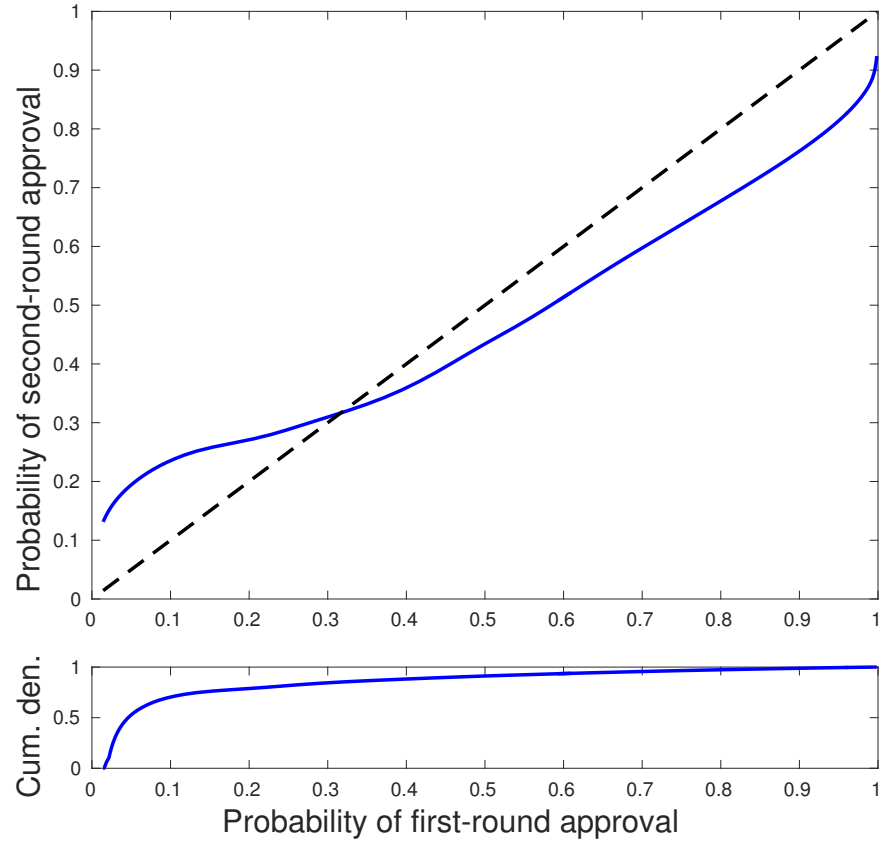
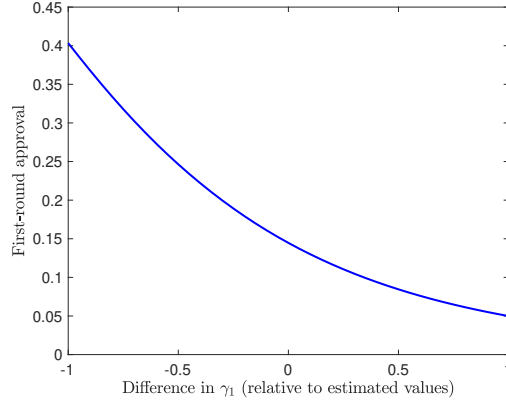


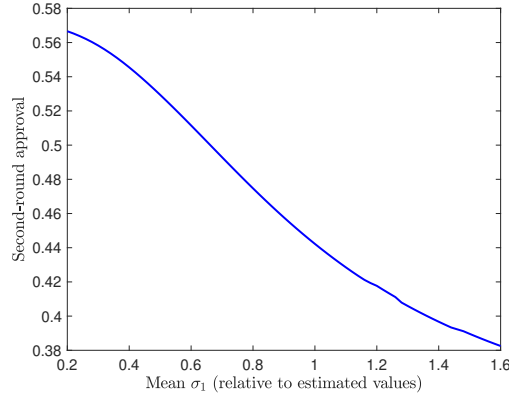
Figure plots first-round approval probability against second-round approval probability conditional on first-round approval for each value of case strength r_i . Secondary graph displays cumulative density of case strength. Black dotted line marks out 45° . See [Section 4.5](#) for full discussion.

Figure 6: Structural parameters and reduced-form moments

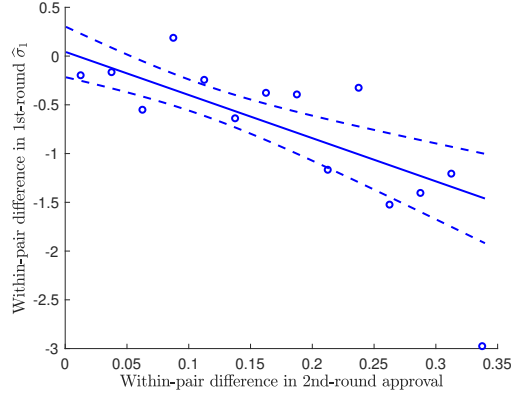
(a) 1st-round approval vs. 1st-round standards γ_1



(b) 2nd-round approval vs. 1st-round error σ_1

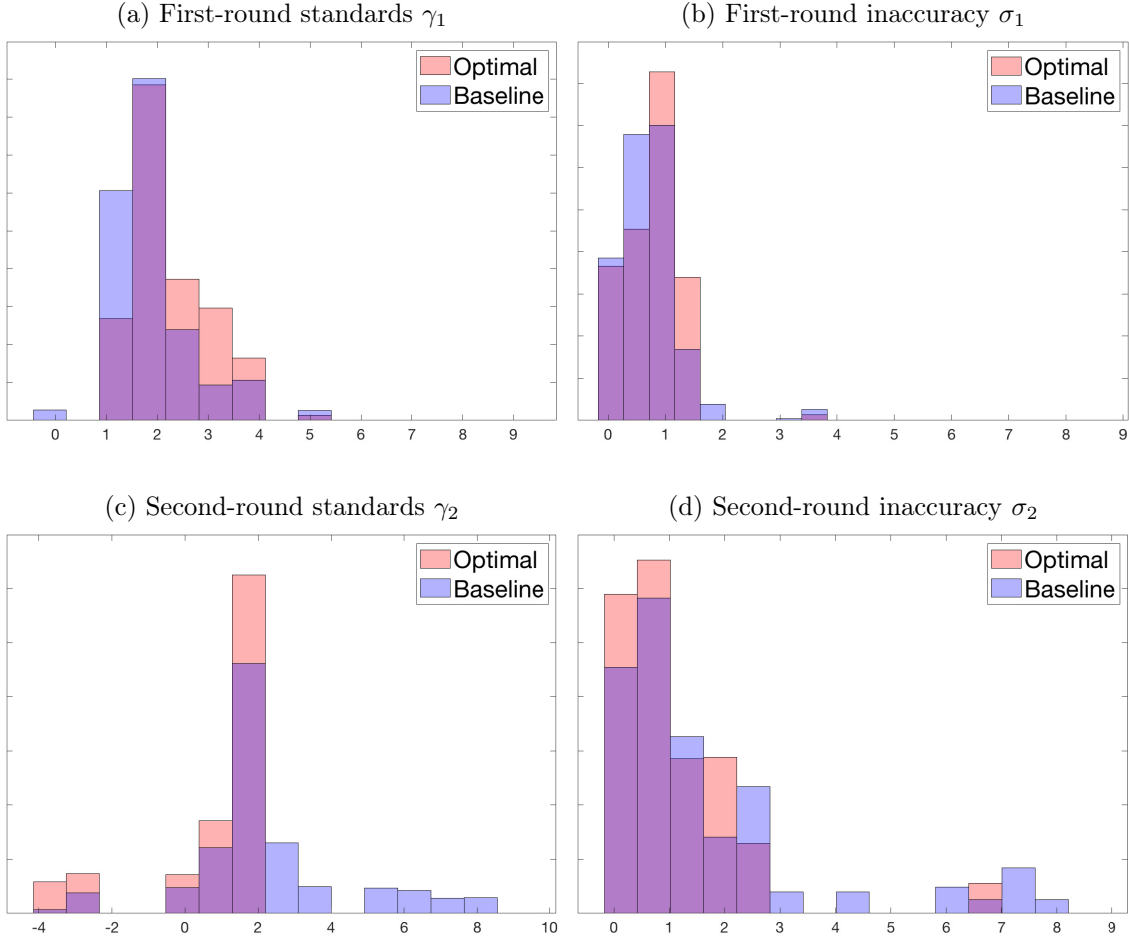


(c) Difference in $\hat{\sigma}_1$ vs. difference in subsequent approval, 1st round judge pairs



This figure shows how reduced-form moments are related to the structural parameters for the decision model $\mathbb{1}[r_i > \gamma_{js} + X_{ijs}\beta_s + \varepsilon_{ijs}]$, $\varepsilon_{ijs} \sim \mathcal{N}(0, \sigma_{js}^2)$. Panel A contains model estimates of the first-round approval probability as a function of shifts from estimated judge thresholds $\hat{\gamma}_{j1}$. Panel B contains model estimates of second-round approval as a function of mean first-round inaccuracy scaled relative to the estimated value—higher accuracy make second-round approval more likely. In Panel C, I match pairs of judges with first-round approval rates within 1 percentage point, then display the difference in estimated inaccuracy $\hat{\sigma}_1$. See [Section 4.6](#) for full discussion.

Figure 7: Distribution of judge parameters under optimal allocation



I minimize judge workload by over the assignment of cases to pairs of judges, under the requirement that (1) no judge works more than she does in the baseline, (2) at least as many claimants are approved, and (3) 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 17.8%. See [Section 4.11](#) for full discussion.

7 Tables

Table 1: Randomization tests of claimant characteristics on judge leniency

	Male	Africa	Asia	South America	IRB mean approval	Predicted approval	1st-round mean approval
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
<i>Panel A: First round judges</i>							
First-round approval rate	0.011 (0.021)	-0.031 (0.030)	-0.089 (0.071)	-0.006 (0.023)	0.010 (0.014)	-0.002 (0.002)	
F-stat	0.88	2.82	9.54	2.47	4.23	2.99	
<i>p</i> -value	0.73	0.00	0.00	0.00	0.00	0.00	
Observations	50,435	50,435	50,435	50,435	50,435	50,435	
<i>Panel B: Second round judges</i>							
Second-round approval rate	-0.048 (0.030)	-0.012 (0.037)	0.000 (0.042)	0.032 (0.042)	-0.005 (0.019)	0.002 (0.002)	-0.024 (0.018)
F-stat	1.01	1.53	1.71	1.22	1.54	2.11	3.07
<i>p</i> -value	0.45	0.01	0.00	0.12	0.00	0.00	0.00
Observations	7,143	7,143	7,143	7,143	7,143	7,143	7,143

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. F-stats come from separate regression of outcome on judge fixed effects. See [Section 4.2](#) for full discussion. 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

	(1)	(2)	(3)
First round judge approval rate	-0.264*** (0.0521)	-0.312*** (0.0423)	-0.324*** (0.0437)
Second round judge approval rate		0.958*** (0.0239)	
Second-round judge FE	No	No	Yes
Observations	8,446	8,446	8,446

See [Section 4.3](#) for full discussion. 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

	Judge-pair round FEs		Judge-pair FEs	
	(1)	(2)	(3)	(4)
Model approval probability	0.924*** (0.130)	0.926*** (0.148)	1.007*** (0.0472)	1.003*** (0.0480)
Model controls	No	Yes	No	Yes
Mean approval	0.44	0.44	0.44	0.44
F-stat for judge pairs	1.01	1.01	0.99	0.99
p -value	0.723	0.718	0.774	0.777
SD of judge-pair EB means	0.001	0.001	0.006	0.006
Observations	8,179	8,179	8,179	8,179

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. See [Section 4.7](#) for full discussion. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: First-round judge accuracy by experience and workload

	(1)	(2)	(3)	(4)
<i>Coefficients ψ_1 affecting judge inaccuracy σ_1</i>				
Experience > 1 year	-0.777*** (0.035)	-1.154*** (0.055)	-0.736*** (0.029)	-0.328*** (0.027)
Experience > 5 years	-0.388*** (0.022)	-0.524*** (0.029)	-0.414*** (0.144)	0.049 (0.048)
Experience > 10 years	-0.411*** (0.026)	-0.747*** (0.050)	-0.762*** (0.073)	-0.936*** (0.032)
Log caseload			0.146*** (0.022)	
Log caseload (≤ 5 yrs exp)				0.305*** (0.005)
Log caseload (> 5 yrs exp)				0.034*** (0.007)
Second-round experience control	Yes	Yes	Yes	Yes
Career number of cases	No	Yes	No	No

Reports coefficients for decision model $\mathbb{1}[r_i > X_{ijs}\beta_s + \gamma_{js} + \tilde{\varepsilon}_{ijs}(W_{ijs})]$, $\tilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, e^{\sigma_{js} + W_{ijs}\psi})$. All models include controls for time/date of decision in β , 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. See [Section 4.8](#) for full discussion. Standard errors clustered at the level of the first round judge. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Inaccuracy before and after judge selection reform

	Baseline			Experience control in σ_1		
	(1)	(2)	(3)	(4)	(5)	(6)
Appointed after reform (=1)	-0.0953 (0.182)	-0.610 (0.370)	-0.663 (0.404)	-1.182*** (0.180)	-1.162*** (0.259)	-1.189*** (0.271)
Liberal appointee (=1)			-0.0105 (0.193)			-0.0237 (0.104)
Male judge (=1)			-0.217 (0.220)			-0.132 (0.129)
Year appointed	No	Yes	Yes	No	Yes	Yes
Pre-reform mean	1.02	1.02	1.02	1.72	1.72	1.72
Number of judges	53	53	53	53	53	53

Estimated with Hanushek (1974) correction for estimated dependent variable. Dependent variable is inaccuracy σ_{j1} , which is estimated from decision model $\mathbb{1}[r_i > X_{ijs}\beta_s + \gamma_{js} + \tilde{\varepsilon}_{ijs}(W_{ijs})]$, $\tilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\tilde{\sigma}_{js} + W_{ijs}\psi_s)$. All models include controls for time/day of decision in β , 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, β_s and ψ_s include dummies for more than 1, 5, and 10 years of experience. See [Section 4.9](#) for full discussion. Standard errors in parentheses and clustered at the judge level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Judge standards γ_1 and inaccuracy σ_1 on lawyer ratings

	γ_1 (mean=1.84, SD=.84)			σ_1 (mean=.57, SD=.45)		
	(1)	(2)	(3)	(4)	(5)	(6)
Lawyer favorability rating, SD	-0.375*		-0.318	-0.076		-0.029
	(0.190)		(0.197)	(0.090)		(0.092)
Lawyer unpredictability rating, SD		0.252**	0.138		0.124**	0.114**
		(0.110)	(0.105)		(0.049)	(0.052)
Observations	73	73	73	73	73	73

Reports linear regressions of model coefficients on lawyer survey responses, estimated with Hanushek (1974) weights for estimated dependent variables. Decision model is $\mathbb{1}[r_i > \gamma_{js} + X_{ijs}\beta_s + \tilde{\varepsilon}_{ijs}(W_{ijs})]$, $\tilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\tilde{\sigma}_{js} + W_{ijs}\psi_s)$. All models include controls for time/day of decision in β , 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 β_s and ψ_s . See [Section 4.10](#) for full discussion. Standard errors clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix for *Examiner Inconsistency: Evidence from Refugee Appeals*

A1 Testing monotonicity

The monotonicity assumption in examiner-assignment IV requires that examiners rank cases in the same way, but differ in the share of cases that they approve (or in other contexts, incarcerate). In this paper I uncover substantive ordering differences between examiners, implying that monotonicity is violated.

In this section I begin by introducing three different monotonicity assumptions, and discussing when each is necessary. Most empirical work has made the first, and most stringent assumption, which I refer to as *pairwise monotonicity*. However, the literature has not tested the strongest implications of pairwise monotonicity, and in [Section A1.2](#) I introduce a test to do so.

Recent work has clarified that pairwise monotonicity is necessary for estimating a nonparametric MTE, but not a LATE, which requires the weaker *average monotonicity* assumption ([Frandsen, Lefgren, and Leslie \(2019\)](#), hereafter FLL). I introduce a new definition of monotonicity that is appropriate for testing the assumption needed for a parametric MTE to recover convex combinations of treatment effects, and show that it nests the pairwise and average cases. In [Section A1.3](#) I introduce a test for the assumption.

A1.1 Three definitions of monotonicity

For individual i assigned to instrument j , define the realized treatment as $T_i(j) \in \{0, 1\}$, and the realized outcome $y_i(j)$. Assume that assignment to one of $J + 1$ examiners is orthogonal to potential outcomes, and order the instruments $j \in \{0, 1, \dots, J\}$ by $p_j = E[T_i(j)]$. The most stringent form of monotonicity requires that assignment to the next-most-severe examiner increases the probability of treatment assignment for each individual. Formally,

Definition 1 (Pairwise monotonicity). *For all individuals i , if $j' > j$ then $T_i(j') \geq T_i(j)$.*

This is the definition from [Angrist, Imbens, and Rubin \(1996\)](#), and along with the relevance, exclusion, and exogeneity assumptions, guarantees that linear IV delivers a convex combination of causal effects for the compliers who are induced into treatment by one of the examiners. For the

individuals induced into treatment by assignment to examiner j rather than $j - 1$, the weight on their causal effect is equal to φ_j .¹

In the presence of defiers who are induced *out* of treatment by assignment to a next-more-severe examiner, the weights on the causal effects can become negative. This can be seen by calculating the weight for each of the 2^{J+1} *compliance types* defined by their response to each consecutive examiner, t_ℓ , $\ell \in \{1, \dots, 2^{J+1}\}$. 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 π_ℓ , and the effect of treatment on each type be θ_ℓ . Then, the weight for type ℓ is

$$\pi_\ell \sum_j \left[\mathbb{1}[t_{\ell,j} = C] - \mathbb{1}[t_{\ell,j} = D] \right] \frac{\varphi_j}{\alpha_j} \quad (\text{A1})$$

where $\alpha_j = p_j - p_{j-1}$. [Equation A1](#) differs from the Imbens and Angrist expression whenever there are defiers. However, many of the compliance types that include defiers still have positive weights, and so IV can still deliver a convex combination of treatment effects. In a contemporaneous working paper, FLL formalize this condition as the requirement that there are only compliance types for whom treatment is positively correlated with mean examiner severity p_j . This motivates a different notion of monotonicity:

Definition 2 (Average monotonicity). *For all individuals i , $\text{cov}(T_i, p_j) \geq 0$ almost surely.*

Average monotonicity is a weaker assumption than pairwise monotonicity. I defer the formal proof to [Appendix A4](#), but the intuition is simple: if high-incarcerating examiners are on average *less* likely to incarcerate a compliance group than their low-incarcerating colleagues—and so the instruments fail average monotonicity—there must be a similar monotonicity violation within at least one of the examiner 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 4 (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 A4.

Since pairwise monotonicity is not required for IV, why do we need it? One important reason is that pairwise monotonicity is necessary for nonparametric estimation of the MTE ([Heckman and Vytlacil, 2005](#)), and so in [Section A1.2](#) I introduce a new test of the assumption.

¹They use λ to denote the IV weights, but to avoid confusion with the monotonicity bias term I substitute φ .

In practice, MTEs are often estimated using polynomial functional forms, which corresponds to assuming that the potential outcomes are polynomials in u , the unobserved first-stage error determining selection into treatment (Heckman et al., 2006). As the degree K of the polynomial increases, the relationship between the outcome and p_j increasingly reflects local variation. For lower values of K non-local changes in outcomes with respect to p_j also affect the estimated MTE at a given value of u . Under this approach, a less stringent condition will ensure the estimated MTE reflects a convex combination of treatment effects:

Definition 3 (Polynomial monotonicity). *For all individuals i , the polynomial first stage $E[T_{ij}|p_{ij}] = \sum_{k=0}^K \psi_k p_{ij}^k$ is monotonically increasing over the support of p almost surely.*

The strength of the polynomial monotonicity assumption depends on K . When $K = 1$, polynomial monotonicity is equivalent to average monotonicity, and when $K = J + 1$, polynomial monotonicity is equivalent to pairwise monotonicity. Intermediate values of K allow more flexible MTEs, at the cost of a more stringent monotonicity assumption.

A1.2 Testing pairwise monotonicity

Like disagreement, monotonicity is an assumption about the joint distribution of treatment assignment under different examiners. This means that the number of j defiers (who would be assigned to treatment under examiner $j - 1$ but not examiner j) is not point-identified, but can be bounded using Fréchet inequalities:

$$\begin{aligned} P[i \text{ is a } j \text{ defier}] &= P[T_i(j) = 0, T_i(j - 1) = 1] \geq \max(0, P[T_i(j - 1) = 1] - P[T_i(j) = 1]) \\ &= \max(0, p_{j-1} - p_j) \\ &= \max(0, -\bar{\alpha}_j) \end{aligned} \tag{A2}$$

where $\bar{\alpha}_j$ comes from a regression of treatment on the dummy z_j indicating assignment to examiner j , among the sample of individuals assigned to examiner j or $j - 1$:

$$T_{ij} = \mu + \bar{\alpha}_j z_j + \varepsilon_{ij} \tag{A3}$$

The right hand side of Equation A2 is necessarily zero, because the examiners have been ordered according to propensity to assign treatment and so $\bar{\alpha}_j \geq 0$. However, Equation A2 also applies conditional on covariates, and so whenever the first stage is negative for a given demographic group, the bound is informative and $Pr[i \text{ is a } j \text{ defier} | D_{ig} = 1] \geq -\alpha_j^g$, where D_{ig} is a dummy variable indicating that individual i belongs to demographic group g , and $\alpha_j^g = E[T_i | j, D_{ig}] - E[T_i | j - 1, D_{ig}]$.

Monotonicity is directly testable by estimating the following equation for individuals assigned to either examiner j or $j - 1$:

$$T_{ijg} = \sum_{g=1}^G \alpha_j^g z_{ij} \times D_{ig} + \mu_g + \varepsilon_{ijg} \quad (\text{A4})$$

The pairwise monotonicity null hypothesis is $H_0 : \alpha_j \geq 0$, where $\alpha_j = [\alpha_j^1, \dots, \alpha_j^G]'$ and the \geq is applied element-wise. This hypothesis can be tested using [Wolak \(1989\)](#), and delivers J tests of pairwise monotonicity. The tests can be combined by estimating

$$T_{ijg} = \sum_{g=1}^G D_g \times \left(\sum_{\ell=1}^J \alpha_\ell^g \mathbb{1}[\ell \leq j] \right) + \mu_g + \varepsilon_{ijg} \quad (\text{A5})$$

Since the examiner indices j are ordered by overall severity, each coefficient α_j^g represents the difference in approval rates for demographic group g between examiner j and examiner $j - 1$. Under pairwise monotonicity, these differences are all positive, so the null hypothesis is $H_0 : \alpha \geq 0$ for $\alpha = [\alpha_1^1, \dots, \alpha_1^G, \dots, \alpha_J^1, \dots, \alpha_J^G]$.

Measuring the distance from pairwise monotonicity

In practice, the test nearly always rejects pairwise monotonicity in judge contexts. A important question is the extent to which monotonicity is violated. One possible answer is motivated by the IV estimand, which is equal to

$$\sum_{j=1}^J \varphi_j (\theta_{C,j} + \lambda_j (\theta_{C,j} - \theta_{D,j})) \quad (\text{A6})$$

where $\theta_{C,j} = \sum_{\ell} \mathbb{1}[t_{\ell,j} = C] \theta_{\ell} \pi_{\ell} / \sum_{\ell} \mathbb{1}[t_{\ell,j} = C] \pi_{\ell}$ is the average treatment effect for j compliers, and $\theta_{D,j} = \sum_{\ell} \mathbb{1}[t_{\ell,j} = D] \theta_{\ell} \pi_{\ell} / \sum_{\ell} \mathbb{1}[t_{\ell,j} = D] \pi_{\ell}$ is the average treatment effect for j defiers. λ_j is the ratio of j defiers to net j compliers ([Angrist, Imbens, and Rubin, 1996](#)), and can be bounded in the following way:

$$\lambda_j = \frac{P(i \text{ is a } j \text{ defier})}{P(i \text{ is a } j \text{ complier}) - P(i \text{ is a } j \text{ defier})} \geq \lambda_j^l = \frac{\sum_g -\alpha_j^g \mathbb{1}[\alpha_j^g < 0] w_j^g}{\bar{\alpha}_j}$$

where w_j^g is the share of the g^{th} demographic group among individuals assigned to examiner j or $j - 1$.

A natural measure of the potential bias in [Equation A6](#) is $\Lambda = \sum_{j=1}^J \varphi_j \lambda_j$, which is analogous to [Angrist et al. \(1996\)](#)'s λ_j measure of potential bias when the instrument is assignment to examiner j rather $j - 1$. Whenever Λ is larger than 0, differences in treatment effects between the complier and

defier groups will bias the LATE estimate away from the standard defined by the Angrist-Imbens-Rubin weights.

The lower bound of Λ , Λ^l , can be estimated by taking the Equation A5 estimates of the first stage, α_j , and jointly maximizing the weighted absolute sum of the negative coefficients:

$$\Lambda = \sum_{j=1}^J \varphi_j \lambda_j \geq \Lambda^l = \sum_{j=1}^J \varphi_j \frac{\sum_g -\alpha_j^g \mathbb{1}[\alpha_j^g < 0] w_j^g}{\bar{\alpha}_j} = \max_{\tau \in \mathcal{T}_{\{-1,0\}^{JG}}} \tau W \tilde{\alpha} \quad (\text{A7})$$

where $\tilde{\alpha} = [\alpha_1^1/\bar{\alpha}_1, \dots, \alpha_1^G/\bar{\alpha}_1, \alpha_2^1/\bar{\alpha}_2, \dots, \alpha_J^G/\bar{\alpha}_J]'$ is the $JG \times 1$ vector of the ratio of demographic-examiner-specific incremental first stages (Equation A5) and examiner incremental first stages (Equation A3),² and W is a $JG \times JG$ weighting matrix defined by the diagonal matrix of $[w_1^1 \varphi_1, \dots, w_1^G \varphi_1, w_2^1 \varphi_2, \dots, w_J^G \varphi_J]$. $\mathcal{T}_{\{-1,0\}^{JG}}$ is the 2^{JG} -size set of all $JG \times 1$ row vectors made up of $\{-1, 0\}$.

This expression is simply the weighted sum of the individual bounds on λ_j . However, the maximization form on the right hand side of Equation A7 is convenient, because it allows the use of the tools in Chernozhukov, Lee, and Rosen (2013) for estimation of half-median unbiased bounds and confidence intervals.

A1.3 Testing average and polynomial monotonicity

While polynomial monotonicity cannot be directly tested, one implication is that the first stage regression of treatment on a polynomial in examiner approval likelihood p_{ij} should be monotonically increasing for all demographic groups. This can be easily tested by estimating the first stage using Bernstein polynomial basis terms. For each group g , define $\tilde{p}_{ijg} = (p_{ijg} - \min(p_{ijg})) / (\max(p_{ijg}) - \min(p_{ijg}))$, and estimate the first stage

$$T_{ijg} = \sum_g \sum_{k=0}^K c_{gk} B_{k,K}(\tilde{p}_{ijg}) + \varepsilon_{ijg} \quad (\text{A8})$$

where $B_{k,K}(\cdot)$ is the k^{th} Bernstein basis polynomial of degree K . Bernstein polynomials are monotonic for group g if and only if $c_{g1} \leq \dots \leq c_{gK}$. Equation A8 can be rewritten as

$$T_{ijg} = \sum_g \sum_{k=0}^K d_{gk} \left\{ \sum_{\ell=k}^K B_{\ell,K}(\tilde{p}_{ijg}) \right\} + \varepsilon_{ijg} \quad (\text{A9})$$

which translates the null hypothesis to $d_{g,1} \geq 0, \dots, d_{g,K} \geq 0$ for all g . Defining $\hat{d} = \{\hat{d}_{gk} | k \geq 1\}$, the Wolak (1989) test statistic is

²These coefficients can be estimated in one regression with $T_{ij} = \mu + \sum_{\ell=1}^J \tilde{\alpha}_\ell \mathbb{1}[\ell \leq j] + \varepsilon_{ij}$.

$$IU = \min_d (\hat{d} - d)' \hat{\Sigma}^{-1} (\hat{d} - d) \quad (\text{A10})$$

$$\text{subject to } d \geq 0 \quad (\text{A11})$$

where $\hat{\Sigma}$ is the estimated covariance matrix of \hat{d} . This form is particularly convenient because it is agnostic to how $\hat{\Sigma}$ is estimated, and so can easily accommodate clustered or robust standard errors.

The distribution of IU is known under the least-favorable null (and is a mixture of chi-squares), so p -values are easy to calculate. In particular,

$$p = \sum_{k=0}^{KG} Pr[\chi_k^2 \geq IU] w(KG, KG - k, \Sigma) \quad (\text{A12})$$

where $w(x, y, z)$ is the likelihood a multivariate normal of dimension x and covariance matrix z will have y positive elements. These weights can be calculated by simulation ([Wolak, 1989](#)).

The most common current approach to testing monotonicity in examiner contexts is to estimate subgroup-specific linear first stages, and check that all the first stages are positive (eg, [Bhuller et al. \(2016\)](#); [Dobbie et al. \(2018\)](#)). The 1-degree polynomial monotonicity test is nearly identical to this approach, but delivers a single p -value rather than one estimate for each subgroup. As the degree of the polynomial increases, the polynomial test becomes more sensitive to local violations of monotonicity. Rejection for low degrees of polynomials suggests that MTEs estimated using the same degree are unlikely to be reliable.

A2 Testing pairwise and polynomial monotonicity

Testing pairwise monotonicity

In [Appendix A1](#) I introduce a new test of one implication of monotonicity: that for each pair of consecutively-more-severe judges, the more lenient judge should approve more 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 highest-approving judge for comparability with the polynomial test, and partial out year and office fixed effects.

I estimate [Equation A5](#) using the demographic groups defined above, and then test the null that all the α are weakly larger than 0. The histogram of t-stats is displayed in [Figure A3](#); although the median coefficient is larger than zero there are many precisely estimated negative coefficients. Unsurprisingly, the inequality test of $\alpha \geq 0$ rejects the null of monotonicity, with a p -value of 1×10^{-10} .

To understand the distance these instruments are from the no-defiers standard, I also estimate bounds on $\Lambda = \sum_{j=1}^J \varphi_j \lambda_j$ using [Equation A7](#). With J judges and G demographic groups, calculating the bound requires checking each of 2^{JG} possible combinations of coefficients in [Equation A5](#), 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 α_j^g (which are main ingredients for the bound) are mutually independent.³ This is true whenever [Equation A5](#) does not contain controls, so I 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 Λ is equal to zero, with a median-unbiased estimate of 1.14 and a 95% confidence interval lower endpoint of 0.91. [Figure A4](#) displays the IV weights, as well as the bound on the IV-weighted λ_j 's for each histogram bin.⁴ 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 λ_j is larger than 0.5, and zero can be rejected.

³Estimation of the CLR bound in this case requires knowing the highest-variance linear combination of the coefficients $\tau\alpha$ where $\tau \in \mathcal{T}_{\{-1,0\}^{JG}}$ is the set of JG length vectors combining -1 and 0. Under independence of α 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.

⁴To be precise, in each block of judges I estimate the lower bound on $\sum_{j \in \text{block}} \varphi_j \lambda_j / (\sum_{k \in \text{block}} \varphi_k)$.

Testing polynomial monotonicity

Table A7 presents tests of polynomial monotonicity. Each row of the table contains the p -values from the inequality test of $\hat{d} \geq 0$ from Equation A9, where the groups are defined as the intersection of the listed characteristics.

In line with other judge contexts, the degree-1 tests do not reject monotonicity for any group definition, with a p -value of 1. For the 3rd degree polynomial, there is a rejection for one group definition. Finally, consistent with the pairwise test, monotonicity is rejected for all groups with 4th and 5th degree polynomials.

These results suggest that the monotonicity violations caused by ordering inaccuracy are mild enough that linear IV will still estimate a convex combination of treatment effects. However, they suggest that the MTE could be more severely biased, which is confirmed by simulations in Appendix A3.

A3 Simulated effect of monotonicity violations on IV estimates

As I show in [Appendix A2](#), in this setting there are large violations of polynomial and pairwise monotonicity, which is reflected by differential harshness across judges towards different claimant groups. The immediate problem could be addressed by interacting the instrument with the demographic groups, potentially selecting interactions with LASSO ([Mueller-Smith, 2015](#)).

However, interacting the instrument and demographic groups will not address monotonicity violations along unobserved dimensions, potentially leading to bias. In this section, I use the model estimates of judge standards and accuracy to explore the effect of inaccuracy on bias in IV and MTE estimates of the effect of first round approval on second round approval.

I begin by calculating the estimated MTE if judges made no ordering errors ($\sigma_{j1} = 0$ for all judges), adjusting standards γ_{j1} to keep each judges approval rate the same. This guarantees that monotonicity is satisfied, and provides a point of comparison against the baseline MTE.

[Figure A5](#) shows the MTE under actual judge behavior, and the perfect-accuracy counterfactual. The bias is relatively large—particularly at the extremes of the support—but is both positive and negative at different points. At the mean judge severity, the baseline MTE is 0.42 while the no-ordering-errors baseline is 0.33.

The direction of bias from ordering errors is *a priori* unknown ([Klein, 2010](#)). In this case, even after observing the estimated MTE, the varying sign of the MTE bias makes it difficult to predict the sign of the IV bias. I calculate it directly, and find that the baseline IV estimate is only slightly biased, from 0.34 under perfect accuracy to 0.36 under the observed level of ordering errors. This concords nicely with the finding in the previous section of pairwise and polynomial monotonicity violations—suggesting bias in MTE estimates—with no rejection of the average monotonicity condition required for a well-defined LATE.

A4 Proof of Theorem 4

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

$$x_{ijg} = \sum_{g=1}^G \alpha_j^g z_j \times D_g + \mu_g + \varepsilon_{ijg} \quad (\text{A13})$$

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 α consists of all judge-demographic combinations of $\alpha_{j,g}$.

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

$$x_{ijg} = \mu_0 + \bar{\alpha}_g \bar{x}_j + \varepsilon_{ijg} \quad (\text{A14})$$

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

I prove Theorem 4 by establishing that $\bar{\alpha}_g$ is a weighted average of α_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 α_j^g must be negative, and so pairwise monotonicity is also violated. However, some α_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 $x_{jg} = x_{0g} + \sum_{k=1}^j \alpha_k^g$. 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:

$$\begin{aligned} E[x_{ijg}\bar{x}_j|g] &= \sum_{j=0}^J \pi_j \bar{x}_j E[x_{ijg}|j, g] \\ &= \sum_{j=0}^J \pi_j \bar{x}_j \left[x_{0g} + \sum_{k=1}^j \alpha_k^g \right] \end{aligned}$$

$$\begin{aligned}
&= \sum_{j=1}^J \pi_j \bar{x}_j \left[\sum_{k=1}^j \alpha_k^g \right] \\
&= \sum_{k=1}^J \alpha_k^g \sum_{j=k}^J \pi_j \bar{x}_j
\end{aligned}$$

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:

$$\bar{\alpha}_g = \sum_{j=1}^J w_j \alpha_j^g, \quad w_j = \frac{\sum_{k=j}^J \pi_k \bar{x}_k}{\sum_{k=1}^J \pi_k \bar{x}_k^2} \quad (\text{A15})$$

where the weights are all positive since \bar{x}_j is demeaned and $\bar{x}_0 < \bar{x}_k$ for all k .

A5 Proof of Theorem 1

Bounds on disagreement can be tightened from the naive Equation 2 bounds using additional information on decisions by subsequent judges. Beginning with the definition of disagreement and expanding using the law of total probability,

$$\delta_{AB}(C) = P[y_{i1}(A) \neq y_{i1}(B)|y_{i2}(C) = 1]P[y_{i2}(C) = 1] + P[y_{i1}(A) \neq y_{i1}(B)|y_{i2}(C) = 0]P[y_{i2}(C) = 0] \quad (\text{A16})$$

Taking the probability of disagreement from the first term on the left side of the expression, $P[y_{i1}(A) \neq y_{i1}(B)|y_{i2}(C) = 1]$, Fréchet inequalities imply that

$$\begin{aligned} & P[y_{i1}(A) \neq y_{i1}(B)|y_{i2}(C) = 1] \\ &= P[y_{i1}(A) = 1, y_{i1}(B) = 0|y_{i2}(C) = 1] + P[y_{i1}(A) = 0, y_{i1}(B) = 1|y_{i2}(C) = 1] \\ &\geq \max\{0, P[y_1|D_1^A, y_2(C) = 1] - P[y_1|D_1^B, y_2(C) = 1]\} + \max\{0, P[y_1|D_1^B, y_2(C) = 1] - P[y_1|D_1^A, y_2(C) = 1]\} \\ &= \max\{P[y_1|D_1^A, y_2(C) = 1] - P[y_1|D_1^B, y_2(C) = 1], P[y_1|D_1^B, y_2(C) = 1] - P[y_1|D_1^A, y_2(C) = 1]\} \end{aligned}$$

where the inequality follows from the Fréchet bound and random assignment of judges to cases. The analogous argument can be made for $P[y_{i1}(A) \neq y_{i1}(B)|y_{i2}(C) = 0]$. Plugging into [Equation A16](#) and expanding the max operators,

$$\begin{aligned} \delta_{AB}(C) \geq \max \bigg\{ & \left[P[y_1|D_1^A, y_2(C) = 1] - P[y_1|D_1^B, y_2(C) = 1] \right] P[y_2(C) = 1] + \left[P[y_1|D_1^A, y_2(C) = 0] - P[y_1|D_1^B, y_2(C) = 0] \right] P[y_2(C) = 0], \\ & \left[P[y_1|D_1^A, y_2(C) = 1] - P[y_1|D_1^B, y_2(C) = 1] \right] P[y_2(C) = 1] + \left[P[y_1|D_1^B, y_2(C) = 0] - P[y_1|D_1^A, y_2(C) = 0] \right] P[y_2(C) = 0], \\ & \left[P[y_1|D_1^B, y_2(C) = 1] - P[y_1|D_1^A, y_2(C) = 1] \right] P[y_2(C) = 1] + \left[P[y_1|D_1^A, y_2(C) = 0] - P[y_1|D_1^B, y_2(C) = 0] \right] P[y_2(C) = 0], \\ & \left[P[y_1|D_1^B, y_2(C) = 1] - P[y_1|D_1^A, y_2(C) = 1] \right] P[y_2(C) = 1] + \left[P[y_1|D_1^B, y_2(C) = 0] - P[y_1|D_1^A, y_2(C) = 0] \right] P[y_2(C) = 0] \bigg\} \quad (\text{A17}) \end{aligned}$$

It is clear that the first argument of the max function in [Equation A17](#) is equal to $P[y_1|D_1^A] - P[y_1|D_1^B]$. The second argument can be expanded to

$$\begin{aligned} & \left[P[y_1|D_1^A, y_2(C) = 1] - P[y_1|D_1^B, y_2(C) = 1] \right] P[y_2(C) = 1] + \left[P[y_1|D_1^B, y_2(C) = 0] - P[y_1|D_1^A, y_2(C) = 0] \right] P[y_2(C) = 0] \\ &= P[y_1|D_1^B] - P[y_1|D_1^A] + 2P[y_2(C) = 1]P[y_1|D_1^A, y_2(C) = 1] - 2P[y_2(C) = 1]P[y_1|D_1^B, y_2(C) = 1] \\ &= P[y_1|D_1^B] - P[y_1|D_1^A] + 2P[y_2(C)|D_1^A, y_1 = 1] \frac{P[y_1|D_1^A]}{P[y_2(C) = 1|D_1^A]} P[y_2(C) = 1] - 2P[y_2(C)|D_1^B, y_1 = 1] \frac{P[y_1|D_1^B]}{P[y_2(C) = 1|D_1^B]} P[y_2(C) = 1] \\ &= P[y_1|D_1^B] - P[y_1|D_1^A] + 2P[y_2(C)|D_1^A, y_1 = 1]P[y_1|D_1^A] - 2P[y_2(C)|D_1^B, y_1 = 1]P[y_1|D_1^B] \end{aligned}$$

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 A17](#) follow analogously, completing the proof.

A6 Judge-assignment identification of second-round accuracy

Identification of second-round accuracy follows a similar principle to first-round accuracy, but since there is no third round to use as a check, identification requires a *non-limiting* judge. Suppose we are trying to determine which second-round judge, A_2 or B_2 , is more accurate. 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 $\tilde{G}_{C_1}(\cdot)/\tilde{G}_{D_1}(\cdot)$ is monotonically increasing wherever $\tilde{G}_{A_2}(\cdot) \neq \tilde{G}_{B_2}(\cdot)$. This is trivially satisfied when $G_{D_1}(\cdot) = 1$ (judge D_1 literally approves everyone), and can be satisfied when judge D_1 is fairly accurate and has low standards γ relative to judge C_1 .⁵

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 the non-limiting judge D_1 ; $\int \tilde{G}_{D_1}(r)\tilde{G}_{A_2}(r)f_r dr = \int \tilde{G}_{D_1}(r)\tilde{G}_{B_2}(r)f_r dr$. 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 accurate than judge B_2 . This can be seen by the following derivation,

$$\begin{aligned}
& \left(P[r > \varepsilon_{A_2} | r > \varepsilon_{C_1}] - P[r > \varepsilon_{B_2} | r > \varepsilon_{C_1}] \right) P[r > \varepsilon_{C_1}] \tag{A18} \\
&= P[r > \varepsilon_{C_1} \cap r > \varepsilon_{A_2}] - P[r > \varepsilon_{C_1} \cap r > \varepsilon_{B_2}] \\
&= \int \tilde{G}_{C_1}(r) [\tilde{G}_{A_2}(r) - \tilde{G}_{B_2}(r)] f_r dr \\
&= \int_{-\infty}^z \frac{\tilde{G}_{C_1}(r)}{\tilde{G}_{D_1}(r)} \tilde{G}_{D_1}(r) [\tilde{G}_{A_2}(r) - \tilde{G}_{B_2}(r)] f_r dr + \int_z^{\infty} \frac{\tilde{G}_{C_1}(r)}{\tilde{G}_{D_1}(r)} \tilde{G}_{D_1}(r) [\tilde{G}_{A_2}(r) - \tilde{G}_{B_2}(r)] f_r dr \\
&> \int_{-\infty}^z \frac{\tilde{G}_{C_1}(z)}{\tilde{G}_{D_1}(z)} \tilde{G}_{D_1}(r) [\tilde{G}_{A_2}(r) - \tilde{G}_{B_2}(r)] f_r dr + \int_z^{\infty} \frac{\tilde{G}_{C_1}(z)}{\tilde{G}_{D_1}(z)} \tilde{G}_{D_1}(r) [\tilde{G}_{A_2}(r) - \tilde{G}_{B_2}(r)] f_r dr \\
&= 0
\end{aligned}$$

where monotonicity of $\tilde{G}_{C_1}(\cdot)/\tilde{G}_{D_1}(\cdot)$ takes the place of monotonicity of second-round approval in the identification of first-round accuracy.

⁵The requirement is also satisfied by assuming that $\tilde{G}_{C_1}(z)/\tilde{G}_{D_1}(z) > \tilde{G}_{C_1}(w)/\tilde{G}_{D_1}(w)$ whenever $z > w$, and $\tilde{G}_{C_1}(z)/\tilde{G}_{D_1}(z) < \tilde{G}_{C_1}(w)/\tilde{G}_{D_1}(w)$ when $z < w$, which is considerably weaker but is harder to understand because it is defined in reference to the point of single-crossing z .

A7 Proof of Theorem 2

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, 2000), where the duration is the number of rounds until a judge rejects the applicant's case (duration is capped at 2). Equation 5 from the main text sets out the problem as identifying the parameters of the choice model where approval in each stage s occurs if

$$r_i > \varepsilon_{ijs}(X_{ijs}, W_{ijs}) = \gamma_{js} + X_{ijs}\beta_s + \tilde{\varepsilon}_{ijs}(W_{ijs}) \quad (\text{A19})$$

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 γ_{js} and β_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 A19, in each stage an individual is approved if

$$\mathbb{1}[-X_{ijs}\beta_s - \gamma_{js} > \tilde{\varepsilon}_{ijs}(W_{ijs}) - r_i] = H_{js,W}(-X_{ijs}\beta_s - \gamma_{js}) \quad (\text{A20})$$

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 $\tilde{\varepsilon}_{ijs}(W_{ijs})$. The assumption of median-zero errors allows nonparametric identification of β_1 and H_{k1} up to scale (Manski, 1975). As in Chen, Heckman, and Vytlacil (2000), I normalize in each round by the norm of β_s , so the identified quantities are $\tilde{\eta}_{k1} = \kappa_1\eta_{k1}$, $\tilde{\beta}_1 = \kappa_1\beta_1$, and $\tilde{\varepsilon}_{k1} = \kappa_1\tilde{\varepsilon}_{k1}$, where $\kappa_s = 1/||\beta_s||$.

Note that 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 γ_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

$$\lim_{X_{ij1}\beta_1 \rightarrow -\infty} \mathbb{1}[-X_{ik2}\beta_2 - \gamma_{k2} > \tilde{\varepsilon}_{ik2}(W_{ik2}) - r_i | -X_{ij1}\beta_1 - \gamma_{j1} > \tilde{\varepsilon}_{ij1}(W_{ij2}) - r_i] = H_{k2,W}(-X_{ij2}\beta_2 - \gamma_{j2}) \quad (\text{A21})$$

so β_2 and H_{k2} are similarly identified to scale after conditioning on values of X_{ij1} so that all first-round claimants are accepted. Simultaneously varying X_1 and X_2 then allows identification of the joint distribution of $\tilde{\eta}_{k1}$ and $\tilde{\eta}_{k2}$.

Decomposing the composite error into permanent and transitory errors is a two step process. I first use the variances and covariance of the composite errors to identify the relative variances. Then, taking those variances as given, I rely on a result from [Rao \(1971\)](#) to decompose the composite errors into their component parts.

Identification of the variance follows from the observation that

$$\begin{aligned} \text{var}(\eta_{k1}) &= \kappa_1^2(1 + \text{var}(\tilde{\varepsilon}_{k1})) \\ \text{var}(\eta_{k2}) &= \kappa_2^2(1 + \text{var}(\tilde{\varepsilon}_{k2})) \\ \text{cov}(\eta_{k1}, \eta_{k2}) &= \kappa_1\kappa_2 \end{aligned}$$

where the normalization of r_i to have a variance of 1 plays a key role. This set of three equations has four unknowns. However, by Assumption 3(d), at least one component of β_1 and β_2 is the same, so the ratio κ_1/κ_2 is known. This leave us with four equations and four unknowns, κ_1 , κ_2 , $\text{var}(\tilde{\varepsilon}_{k1})$ and $\text{var}(\tilde{\varepsilon}_{k2})$, so all are identified.

The goal is now to identify the distributions of r_i and $\tilde{\varepsilon}_{ks}$ from $\eta_{ks} = \kappa_s r + \kappa_s \tilde{\varepsilon}_{ks}$, with κ_s and the variance of r known. Note that η_{k1} and η_{k2} are linear combinations of the independent random variables r_i , $\tilde{\varepsilon}_{k1}$ and $\tilde{\varepsilon}_{k2}$, with known weights. [Rao \(1971\)](#) shows that observing p of these linear combinations identifies up to $p(p+1)/2$ components. Here, we observe the first and second round so $p = 2$ and we can identify each of the three components, the distributions $G_{j1,W}$, $G_{j2,W}$, and F_r .

A8 Estimation details for structural model

For notational simplicity, I collapse all coefficients and regressors into the distribution of the error ε_s , which I denote with mean μ_s and standard deviation σ_s . I first explain the derivation of first-round approval probabilities, then the second-round probabilities.

A8.1 First round approval

$$\begin{aligned} P(r - \varepsilon_1 > 0) &= \int_0^\infty P(r > \tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_1) \\ &= \int_0^{x_m} P(r > \tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_1) + \int_{x_m}^\infty P(r > \tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_1) \end{aligned} \quad (\text{A22})$$

The first term in Equation A22 can be shown to be equal to $\Phi\left[\frac{\ln(x_m) - \mu_1}{\sigma_1}\right]$. Then,

$$\begin{aligned} \int_{x_m}^\infty P(r > \tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_1) &= \int_{x_m}^\infty \frac{x_m^\alpha}{\tilde{\varepsilon}_1^\alpha} \phi\left(\frac{\ln(\tilde{\varepsilon}_1) - \mu_1}{\sigma_1}\right) \frac{1}{\sigma_1 \tilde{\varepsilon}_1} d\tilde{\varepsilon}_1 \\ &= x_m^\alpha \int_{\frac{\ln(x_m) - \mu_1}{\sigma_1}}^\infty e^{-\alpha(\sigma_1 y + \mu_1)} \phi(y) dy \\ &= x_m^\alpha e^{-\alpha\mu_1} \int_{\frac{\ln(x_m) - \mu_1}{\sigma_1}}^\infty e^{-\alpha\sigma_1 y - \frac{y^2}{2}} dy \\ &= x_m^\alpha e^{-\alpha\mu_1 + \frac{\alpha^2\sigma_1^2}{2}} \int_{\frac{\ln(x_m) - \mu_1}{\sigma_1}}^\infty e^{-\frac{1}{2}(y + \alpha\sigma_1)^2} dy \\ &= x_m^\alpha e^{-\alpha\mu_1 + \frac{\alpha^2\sigma_1^2}{2}} \int_{\frac{\ln(x_m) - \mu_1}{\sigma_1} + \alpha\sigma_1}^\infty e^{-\frac{1}{2}y^2} dy \\ &= x_m^\alpha e^{-\alpha\mu_1 + \frac{\alpha^2\sigma_1^2}{2}} \left[1 - \Phi\left(\frac{\ln(x_m) - \mu_1}{\sigma_1} + \alpha\sigma_1\right) \right] \end{aligned} \quad (\text{A23})$$

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; $-\frac{1}{2}(y^2 + 2\alpha\sigma_1 y) = -\frac{1}{2}(y^2 + 2\alpha\sigma_1 y + \alpha^2\sigma_1^2) + \frac{\alpha^2\sigma_1^2}{2} = -\frac{1}{2}(y + \alpha\sigma_1)^2 + \frac{\alpha^2\sigma_1^2}{2}$.

A8.2 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

$$P(r > \varepsilon_2 \cap r > \varepsilon_1) = \int_0^\infty \int_0^\infty P(r > \tilde{\varepsilon}_2 | r > \tilde{\varepsilon}_1) P(r > \tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_2) \quad (\text{A24})$$

The terms inside the integrals can be rewritten

$$P(r > \tilde{\varepsilon}_1) = \mathbb{1}[\tilde{\varepsilon}_1 < x_m] + \mathbb{1}[\tilde{\varepsilon}_1 \geq x_m] \frac{x_m^\alpha}{\tilde{\varepsilon}_1^\alpha} \quad (\text{A25})$$

and

$$\begin{aligned} P(r > \tilde{\varepsilon}_2 | r > \tilde{\varepsilon}_1) = & \mathbb{1}[\tilde{\varepsilon}_1 < x_m] \left[\mathbb{1}[\tilde{\varepsilon}_2 < x_m] + \mathbb{1}[\tilde{\varepsilon}_2 \geq x_m] \frac{x_m^\alpha}{\tilde{\varepsilon}_2^\alpha} \right] + \\ & \mathbb{1}[\tilde{\varepsilon}_1 \geq x_m] \left[\mathbb{1}[\tilde{\varepsilon}_2 < \tilde{\varepsilon}_1] + \mathbb{1}[\tilde{\varepsilon}_2 \geq \tilde{\varepsilon}_1] \frac{\tilde{\varepsilon}_1^\alpha}{\tilde{\varepsilon}_2^\alpha} \right] \end{aligned} \quad (\text{A26})$$

Substituting into Equation A24 and expanding the integrals,

$$\begin{aligned} P(r > \varepsilon_2 \cap r > \varepsilon_1) = & \int_0^\infty \int_0^{x_m} \mathbb{1}[\tilde{\varepsilon}_2 < x_m] + \mathbb{1}[\tilde{\varepsilon}_2 \geq x_m] \frac{x_m^\alpha}{\tilde{\varepsilon}_2^\alpha} dF(\tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_2) \\ & + \int_0^\infty \int_{x_m}^\infty \frac{x_m^\alpha}{\tilde{\varepsilon}_1^\alpha} \left[\mathbb{1}[\tilde{\varepsilon}_2 < \tilde{\varepsilon}_1] + \mathbb{1}[\tilde{\varepsilon}_2 \geq \tilde{\varepsilon}_1] \frac{\tilde{\varepsilon}_1^\alpha}{\tilde{\varepsilon}_2^\alpha} \right] dF(\tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_2) \end{aligned}$$

Further separate the integrals into four components:

$$\int_0^{x_m} \int_0^{x_m} dF(\tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_2) \quad (\text{A27})$$

$$\int_{x_m}^\infty \int_0^{x_m} \frac{x_m^\alpha}{\tilde{\varepsilon}_2^\alpha} dF(\tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_2) \quad (\text{A28})$$

$$\int_{x_m}^{\infty} \int_0^{\tilde{\varepsilon}_1} \frac{x_m^\alpha}{\tilde{\varepsilon}_1^\alpha} dF(\tilde{\varepsilon}_2) dF(\tilde{\varepsilon}_1) \quad (\text{A29})$$

$$\int_{x_m}^{\infty} \int_{\tilde{\varepsilon}_1}^{\infty} \frac{x_m^\alpha}{\tilde{\varepsilon}_2^\alpha} dF(\tilde{\varepsilon}_2) dF(\tilde{\varepsilon}_1) \quad (\text{A30})$$

These four equations (A27-A30) 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:

$$\int_0^{x_m} \int_0^{x_m} dF(\tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_2) = \Phi\left(\frac{x_m - \mu_1}{\sigma_1}\right) \Phi\left(\frac{x_m - \mu_2}{\sigma_2}\right)$$

$$\begin{aligned} \int_{x_m}^{\infty} \int_0^{x_m} \frac{x_m^\alpha}{\tilde{\varepsilon}_2^\alpha} dF(\tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_2) &= x_m^\alpha \Phi\left(\frac{x_m - \mu_1}{\sigma_1}\right) \int_{x_m}^{\infty} e^{-\alpha \ln \tilde{\varepsilon}_2} dF(\tilde{\varepsilon}_2) \\ &= x_m^\alpha \Phi\left(\frac{x_m - \mu_1}{\sigma_1}\right) e^{-\alpha \mu_2 + \frac{\alpha^2 \sigma_2^2}{2}} \left[1 - \Phi\left(\frac{\ln(x_m) - \mu_2}{\sigma_2} + \alpha \sigma_2\right)\right] \end{aligned}$$

The last two make use of the additional fact that

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

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

$$\begin{aligned} \int_{x_m}^{\infty} \int_0^{\tilde{\varepsilon}_1} \frac{x_m^\alpha}{\tilde{\varepsilon}_1^\alpha} dF(\tilde{\varepsilon}_2) dF(\tilde{\varepsilon}_1) &= \int_{x_m}^{\infty} \frac{x_m^\alpha}{\tilde{\varepsilon}_1^\alpha} \Phi\left(\frac{\ln(\tilde{\varepsilon}_1) - \mu_2}{\sigma_2}\right) dF(\tilde{\varepsilon}_1) \\ &= x_m^\alpha \int_{x_m}^{\infty} e^{-\alpha \ln(\tilde{\varepsilon}_1)} \Phi\left(\frac{\ln(\tilde{\varepsilon}_1) - \mu_2}{\sigma_2}\right) \phi\left(\frac{\ln(\tilde{\varepsilon}_1) - \mu_1}{\sigma_1}\right) \frac{1}{\sigma_1 \tilde{\varepsilon}_1} d\tilde{\varepsilon}_1 \end{aligned}$$

$$\begin{aligned}
&= x_m^\alpha e^{-\alpha\mu_1} \int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}}^{\infty} e^{-\alpha\sigma_1 y} \Phi\left(\frac{\sigma_1 y + \mu_1 - \mu_2}{\sigma_2}\right) \phi(y) dy \\
&= x_m^\alpha e^{-\alpha\mu_1 + \frac{\alpha^2\sigma_1^2}{2}} \int_{\frac{\ln(x_m)-\mu_1}{\sigma_1} + \alpha\sigma_1}^{\infty} \Phi\left(\frac{\sigma_1 y - \alpha\sigma_1^2 + \mu_1 - \mu_2}{\sigma_2}\right) \phi(y) dy \\
&= x_m^\alpha e^{-\alpha\mu_1 + \frac{\alpha^2\sigma_1^2}{2}} \int_{\frac{\ln(x_m)-\mu_1}{\sigma_1} + \alpha\sigma_1}^{\infty} \Phi\left(\frac{y - \alpha\sigma_1 + (\mu_1 - \mu_2)/\sigma_1}{\sigma_2/\sigma_2}\right) \phi(y) dy \\
&= x_m^\alpha e^{-\alpha\mu_1 + \frac{\alpha^2\sigma_1^2}{2}} BvN\left(\frac{(\mu_1 - \mu_2)/\sigma_1 - \alpha\sigma_1}{\sqrt{\sigma_2^2/\sigma_1^2 + 1}}, -\frac{\ln(x_m) - \mu_1}{\sigma_1} - \alpha\sigma_1, \frac{1}{\sqrt{\sigma_2^2/\sigma_1^2 + 1}}\right)
\end{aligned}$$

$$\begin{aligned}
\int_{x_m}^{\infty} \int_{\tilde{\varepsilon}_1}^{\infty} \frac{x_m^\alpha}{\tilde{\varepsilon}_2^\alpha} dF_1 dF_2 &= \int_{x_m}^{\infty} x_m^\alpha e^{-\alpha\mu_2 + \frac{\alpha^2\sigma_2^2}{2}} \left[1 - \Phi\left(\frac{\ln(\tilde{\varepsilon}_1) - \mu_2}{\sigma_2} + \alpha\sigma_1\right)\right] dF(\tilde{\varepsilon}_1) \\
&= \tilde{B} \left\{ \left[1 - \Phi\left(\frac{\ln(x_m) - \mu_1}{\sigma_1}\right)\right] - \int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}}^{\infty} \Phi\left(\frac{y + (\mu_1 - \mu_2 + \alpha\sigma_2^2)/\sigma_1}{\sigma_2/\sigma_1}\right) \phi(y) dy \right\} \\
&= \tilde{B} \left\{ \left[1 - \Phi\left(\frac{\ln(x_m) - \mu_1}{\sigma_1}\right)\right] - BvN\left(\frac{(\mu_1 - \mu_2 + \alpha\sigma_2^2)/\sigma_1}{\sqrt{\sigma_2^2/\sigma_1^2 + 1}}, -\frac{\ln(x_m) - \mu_1}{\sigma_1}, \frac{1}{\sqrt{\sigma_2^2/\sigma_1^2 + 1}}\right) \right\}
\end{aligned}$$

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

A8.3 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 A30 in detail; the same method works for all the joint first- and second-round probabilities.

$$\begin{aligned}
&\int_{x_m}^{\infty} \int_{\tilde{\varepsilon}_1}^{\infty} \frac{x_m^\alpha}{\tilde{\varepsilon}_2^\alpha} dF_2 dF_1 \\
&= \int_{x_m}^{\infty} \int_{\tilde{\varepsilon}_1}^{\infty} \frac{x_m^\alpha}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}\varepsilon_1\varepsilon_2} e^{-\alpha\ln(\varepsilon_2) - \frac{1}{2(1-\rho^2)}\left(\left(\frac{\ln(\varepsilon_1)-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{\ln(\varepsilon_1)-\mu_1}{\sigma_1}\right)\left(\frac{\ln(\varepsilon_2)-\mu_2}{\sigma_2}\right) + \left(\frac{\ln(\varepsilon_2)-\mu_2}{\sigma_2}\right)^2\right)} d\varepsilon_2 d\varepsilon_1 \\
&= x_m^\alpha e^{-\alpha\mu_2} \int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}}^{\infty} \int_{\frac{\sigma_1 y + \mu_1 - \mu_2}{\sigma_2}}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} e^{\frac{-1}{2(1-\rho^2)}(2\alpha\sigma_2(1-p^2)x+y^2-2\rho yx+x^2)} dx dy
\end{aligned}$$

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

$$x_m^\alpha e^{-\alpha\mu_2 + \frac{\alpha^2\sigma_2^2}{2}} \int_{\frac{\ln(x_m) - \mu_1}{\sigma_1}}^{\infty} \left(1 - \Phi\left(\frac{\frac{\sigma_1 y + \mu_1 - \mu_2}{\sigma_2} - (\rho y - \alpha\sigma_2(1 - \rho^2))}{\sqrt{1 - \rho^2}}\right) \right) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y + \alpha\sigma_2\rho)^2} dy$$

Rearrange the term in the normal:

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

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.

$$\begin{aligned} &= x_m^\alpha e^{-\alpha\mu_2 + \frac{\alpha^2\sigma_2^2}{2}} \int_{\frac{\ln(x_m) - \mu_1}{\sigma_1} + \alpha\sigma_2\rho}^{\infty} \left(1 - \Phi\left(\frac{y - \alpha\sigma_2\rho + (\mu_1 - \mu_2 + \alpha\sigma_2^2(1 - \rho^2))/(\sigma_1 - \rho\sigma_2)}{\sigma_2\sqrt{1 - \rho^2}/(\sigma_1 - \rho\sigma_2)}\right) \right) \phi(y) dy \\ &= \tilde{B} \left\{ \left[1 - \Phi\left(\frac{\ln(x_m) - \mu_1}{\sigma_1} + \alpha\sigma_2\rho\right) \right] - BvN\left(\frac{-b}{\sqrt{a^2 + 1}}, -\frac{\ln(x_m) - \mu_1}{\sigma_1} - \alpha\sigma_2\rho, \frac{1}{\sqrt{a^2 + 1}}\right) \right\} \end{aligned}$$

$$\tilde{B} = x_m^\alpha e^{-\alpha\mu_2 + \frac{\alpha^2\sigma_2^2}{2}}$$

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

$$a = \sigma_2\sqrt{1 - \rho^2}/(\sigma_1 - \rho\sigma_2)$$

A9 Model assumption checks and robustness

A9.1 Decision timing as regressors

Identification requires that the case timing regressors affect judge thresholds γ_{js} but are not correlated with judge errors $\tilde{\varepsilon}_{ijs}$ 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 A5](#). 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_{ijs} sizably affects judge thresholds γ_{js} , and that the regressors make a substantive contribution to identification.

Another fear is that decision timing affects approval through changing judge accuracy σ_{js} rather than judge standards γ_{js} . I test this directly in [Table A6](#), where I include in turn the two decision timing regressors in W_{ijs} . Since the noon-hearing regressor affects only second-round outcomes these should be interpreted as tests on second-round identification.

In line with the relevance tests in [Table A5](#), both regressors have strong and statistically significant effects on γ_{js} . However, while both are statistically significant, the effect of the regressors on judge accuracy is small. The average marginal effects of the regressors on first-round approval probability through the β_1 channel is 7 and 2.9 times larger than through the ψ_1 channel for the lunch-hearing and end-of-week regressor, respectively.

A9.2 Reduced-form and structural parameters for second-round accuracy

Identification of second-round accuracy relies on matching pairs of second-round judges with similar approval rates conditional on first-round approval by a very lenient, *non-limiting* first-round judge. [Equation A18](#) shows that second-round approval rates conditional on first-round approval by a different, less lenient judge will be higher for the more accurate second-round judge. [Figure A1](#) 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 non-limiting 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 inaccuracy σ_{j2} and the difference in second-round approval rates conditional on first-round approval by all other judges, and shows that the larger the difference in approval rates, the larger the difference in estimated σ_{j2} . Higher approval rates under the comparison judges correspond to improved accuracy (lower σ_{j2}).

A9.3 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 the main results from the paper using estimates from a model identified without regressors. 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 regressors X_{ijs} indeed help with identification.

In Section 4.7 I presented evidence that judicial ordering errors are idiosyncratic observational errors, rather than permanent differences between judges in tastes. 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 A8](#) 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 0.007 in the both models—and the F-test similarly does not reject that the judge pair effects are jointly zero.

In [Table A9](#), 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 (Column 3), though only for judges with fewer than 6 years of experience (Column 4).

Table A10 contains estimates of the effect of judicial selection reform on judge accuracy. As I describe in Section 1.3, the reform made it much more difficult for the government to appoint unqualified judges after 1988. In Section 4.9 I show that this increased baseline estimates of accuracy by approximately 70%. In the model estimated without regressors, the results are also large, declining by 1.2 from a pre-reform mean of 1.6.

Finally, Table A11 mirrors the results of the baseline model with respect to the relationship between the judge decision parameters and the survey of lawyers. Judge standards γ_{js} are negatively correlated with survey measures of judge favorability to claimants, and model-estimated inaccuracy σ_{j1} is negatively correlated with surveyed predictability. Again, this suggests that the correlation between model and survey results are not driven by the use of regressors in identification.

A10 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 sample judges. I was able to find online contact information for 551 of them.⁶ 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, 64 lawyers responded for an overall response rate of 14%.⁷ Table A12 compares responders to non-responders and lawyers for whom I couldn't find contact information. The main difference is 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 actually had experience with. 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. On a scale from 1 to 5, how would you rate the listed judges in terms of **consistency**? Are

⁶The 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.

⁷This 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).

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.

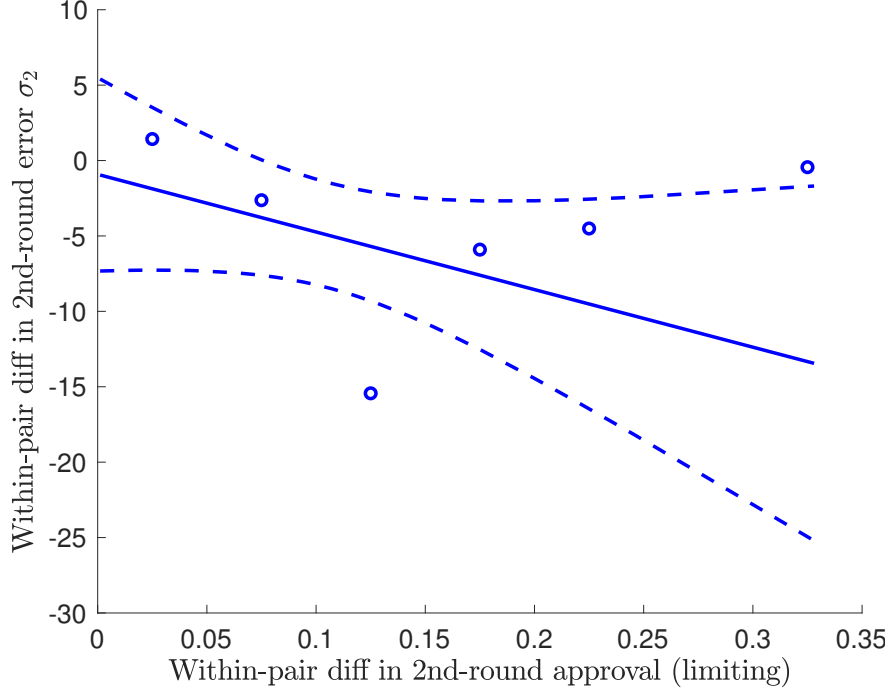
I expected that the first question would be related to judge standards γ_j , and the second with the size of the ordering error σ_j .

Each response was on a five-point likert scale (I reverse the ordering of the accuracy response so it is analogous to the estimated inaccuracy coefficients). I normalize responses by the mean and standard deviation.

I discuss the main results in Section 4.10, where I include only the first two questions. The final question of the survey, which asked about the judges accuracy (using the word in a more normative sense than I have in the rest of his paper), 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 the question into the model, since it implies a normative judgement about the correct outcome of the case. 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 A13](#) adds the last response to the regression of model coefficients on survey responses; the relationship between favorability and γ_2 , and accuracy and σ_1 is almost unchanged.

A11 Appendix Figures

Figure A1: Structural and reduced form measures of second-round accuracy



This figure relates to the estimated decision model $\mathbb{1}[r_i > \gamma_{js} + X_{ijs}\beta_s + \tilde{\varepsilon}_{ijs}]$, $\tilde{\varepsilon}_{ijs} \sim \mathcal{N}(0, \sigma_{js}^2)$. 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 σ_2 as a function of within-pair differences in second-round approval rates conditional on first-round approval by all other judges. See [Appendix A6](#) for more details.

Figure A2: Distribution of judge parameters, no-regressor model

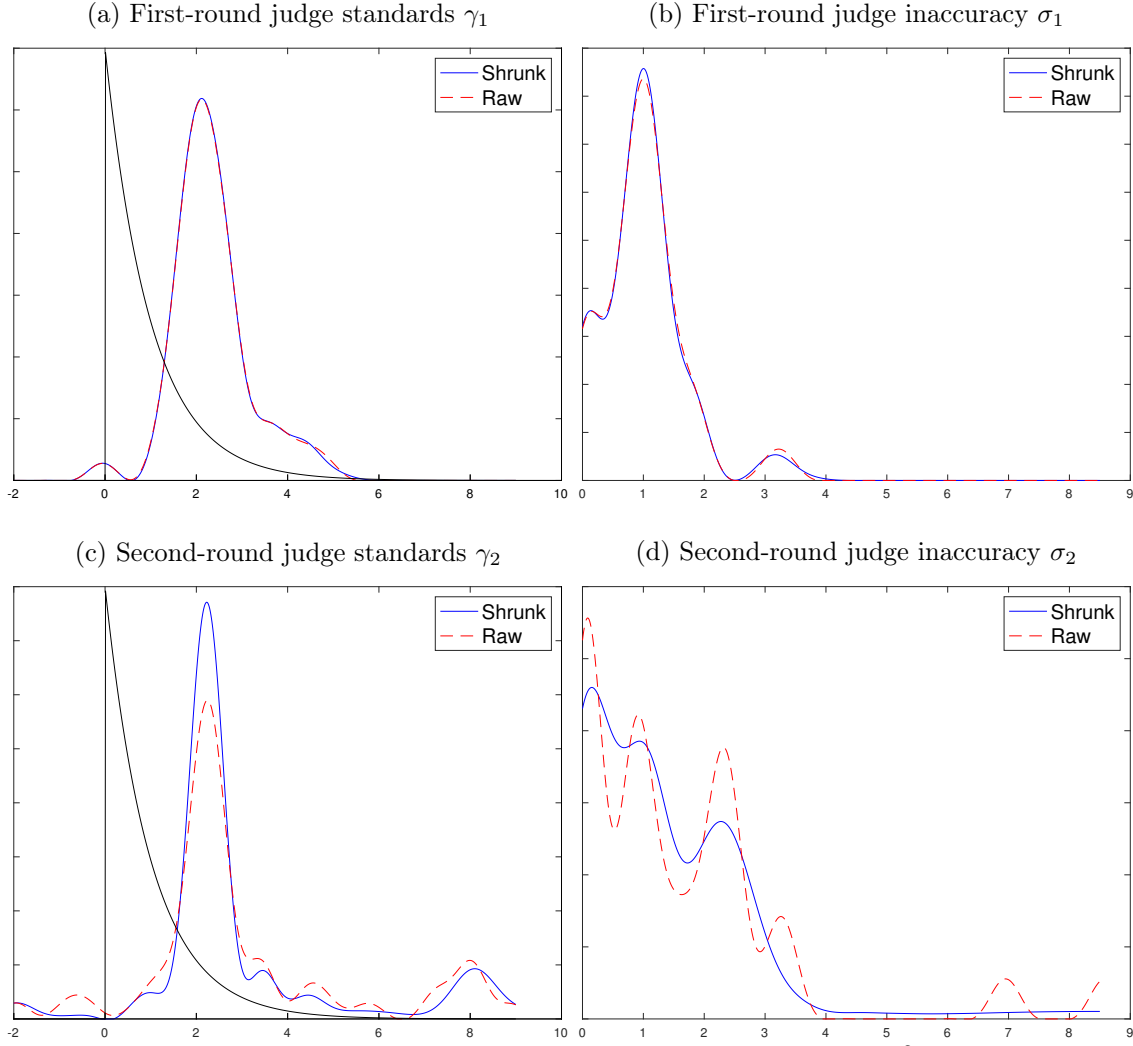
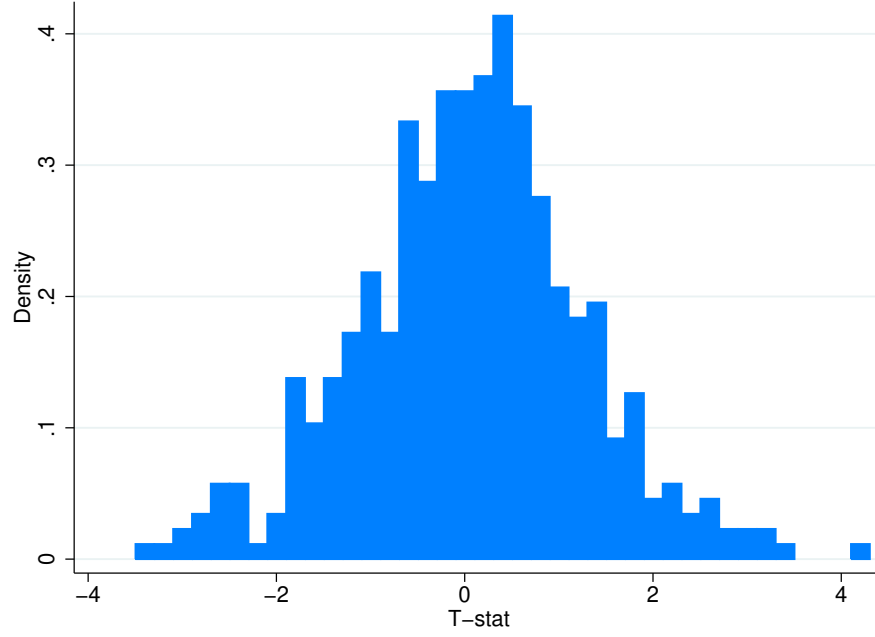


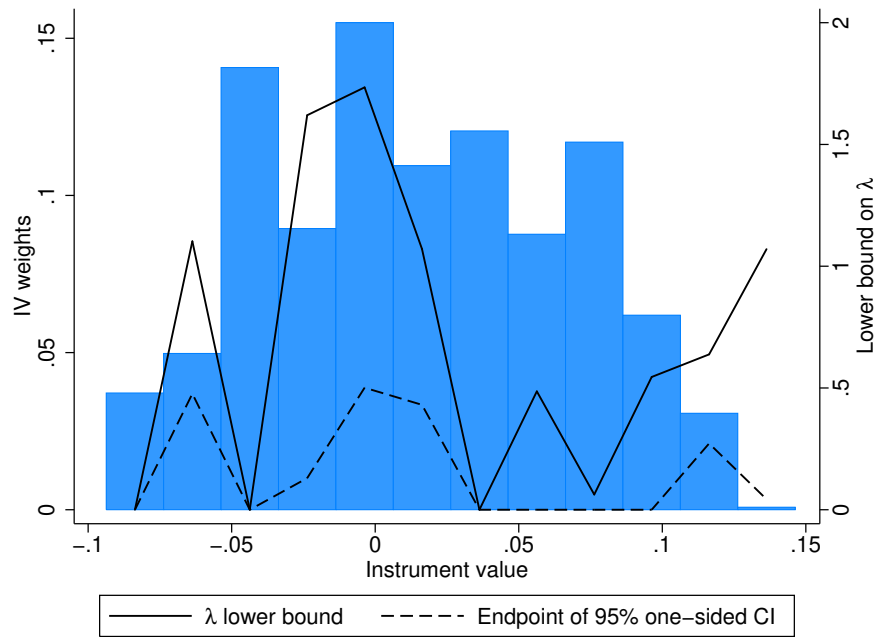
Figure displays coefficients for decision model $\mathbb{1}[r_i > \gamma_{js} + \tilde{\varepsilon}_{ijs}]$, $\tilde{\varepsilon}_{ijs} \sim \mathcal{N}(0, \sigma_{js}^2)$. All models allow the parameters of the Pareto distribution of r_i to vary flexibly by office of origination as well as after 2002 relative to before 2002, when there were legislative changes that may have impacted the distribution of case quality. In contrast to the baseline model reported in Figure 4, this model does not use timing regressors for identification. Each panel contains the density of the raw and shrunk estimates of the judge thresholds γ_1 and γ_2 , and judge inaccuracy σ_1 and σ_2 . Black line is density of case quality r_i . Shrunk estimates recovered via deconvolution of estimates accounting for coefficient-specific standard errors (Delaigle, Hall, and Meister, 2008).

Figure A3: T-stats for pairwise judge-demographic effects



The figure displays a histogram of the t-stats for α_j^g from Equation A5, 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 A4: Bounds on IV-weighted judge-pair λ_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 λ_j 's for judges of that severity. The dashed line marks the one-sided 95% confidence interval.

Figure A5: Estimated MTE at baseline and with no monotonicity violations

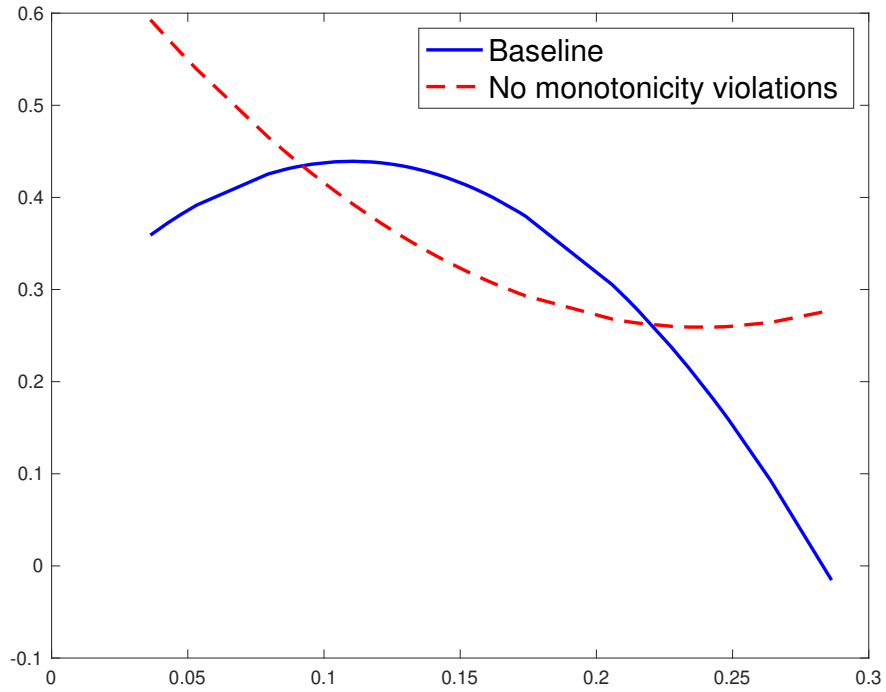


Figure demonstrates how monotonicity violations caused by cross-judge differences in case ordering affects the estimated MTE. I plot the baseline MTE in solid blue, then use model estimates to construct an estimate of the MTE if all judges satisfied monotonicity by being perfectly accurate (ie, $\sigma_{j1} = 0$). MTE estimated by parametric regression of simulated outcome on a cubic polynomial of judge probability of approval, then taking the derivative. Linear IV estimate is 0.36 in the baseline, versus 0.34 without monotonicity violations.

A12 Appendix Tables

Table A1: Judge and case summary statistics

	Mean	SD	Min	Max
Male judge (=1)	0.75	0.44	0.00	1.00
Liberal appointee (=1)	0.72	0.45	0.00	1.00
Experience (years)	6.51	5.63	0.00	28.00
Workload	-0.07	0.80	-3.45	1.53
Male (=1)	0.63	0.43	0.00	1.00
African (=1)	0.19	0.39	0.00	1.00
Asia (=1)	0.10	0.31	0.00	1.00
South American (=1)	0.35	0.48	0.00	1.00
Calgary (=1)	0.02	0.14	0.00	1.00
Montreal (=1)	0.42	0.49	0.00	1.00
Ottawa (=1)	0.02	0.13	0.00	1.00
Vancouver (=1)	0.03	0.18	0.00	1.00
Observations	58,604			

Table A2: Second-round approval on model approval probability and judge characteristics

	French	Liberal	Male	Pre-reform	Exp>1	Exp>5
	(1)	(2)	(3)	(4)	(5)	(6)
Model approval probability	0.998*** (0.0320)	1.004*** (0.0294)	1.009*** (0.0291)	1.005*** (0.0299)	1.001*** (0.0288)	1.002*** (0.0289)
Characteristic, first-round judge (=1)	-0.0116 (0.0146)	-0.0325 (0.0217)	-0.0111 (0.0272)	0.00640 (0.0186)	-0.0124 (0.0416)	-0.0159 (0.0142)
Characteristic, second-round judge (=1)	-0.00702 (0.0157)	-0.0414* (0.0218)	-0.0309 (0.0265)	-0.00433 (0.0166)	-0.0924** (0.0438)	-0.0119 (0.0148)
Characteristic, both judges (=1)	0.00965 (0.0215)	0.0571** (0.0258)	0.0350 (0.0301)	0.0292 (0.0377)	0.0881* (0.0468)	0.0273 (0.0205)
Model controls	Yes	Yes	Yes	Yes	Yes	Yes
Mean approval	0.44	0.44	0.44	0.44	0.44	0.44
F-stat for characteristic pairs	0.26	1.78	1.19	0.44	4.37	0.63
<i>p</i> -value	0.875	0.146	0.340	0.714	0.008	0.585
SD of judge-pair EB means	0.044	0.003	0.001	0.002	0.007	0.003
Observations	8,179	8,179	8,179	8,179	8,179	8,179

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 A3: Approval rate for first-round judges before and after reform

	Approval rate			Approval, year residualized		
	(1)	(2)	(3)	(4)	(5)	(6)
Appointed after reform (=1)	0.00896 (0.0318)	0.0764 (0.0716)	0.0734 (0.0757)	0.0103 (0.0309)	0.0891 (0.0716)	0.0831 (0.0751)
Liberal appointee (=1)			0.00317 (0.0192)			0.0102 (0.0194)
Male judge (=1)			-0.00876 (0.0257)			-0.0140 (0.0250)
Year appointed	No	Yes	Yes	No	Yes	Yes
Pre-reform mean	0.15	0.15	0.15	0.00	0.00	0.00
Number of judges	53	53	53	53	53	53

Regression and dependent variable mean estimated with Hanushek (1974) weights for estimated dependent variable. Robust standard errors in parentheses and clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Inaccuracy σ_{j1} before and after reform, controlling for approval rate

	Baseline			Experience control in σ_1		
	(1)	(2)	(3)	(4)	(5)	(6)
Appointed after reform (=1)	-0.0952 (0.182)	-0.630* (0.354)	-0.680* (0.383)	-1.177*** (0.179)	-1.118*** (0.268)	-1.144*** (0.279)
Liberal appointee (=1)			-0.0112 (0.195)			-0.0219 (0.104)
Male judge (=1)			-0.215 (0.225)			-0.137 (0.133)
Year appointed	No	Yes	Yes	No	Yes	Yes
Pre-reform mean	1.02	1.02	1.02	1.72	1.72	1.72
N umber of judges	53	53	53	53	53	53

Estimated with Hanushek (1974) weights for estimated dependent variable and a control for judge approval rate. Dependent variable is consistency σ_{j1} , which is estimated from decision model $\mathbb{1}[r_i > X_{ijs}\beta_s + \gamma_{js} + \tilde{\varepsilon}_{ijs}(W_{ijs})]$, $\tilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\tilde{\sigma}_{js} + W_{ijs}\psi_s)$. All models include controls for time/day of decision in β , 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, β_s and ψ_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 A5: Placebo and relevance tests for regressors

	Predicted approval		Actual approval	
	(1)	(2)	(3)	(4)
<i>Panel A: First round</i>				
End of week	0.000 (0.000)	0.000 (0.000)	-0.008*** (0.002)	-0.007*** (0.002)
Observations	58,604	58,604	58,604	58,604
<i>Panel B: Second round</i>				
End of week	0.001 (0.001)	0.001 (0.001)	-0.022* (0.012)	-0.022* (0.012)
Noon hearing	-0.001 (0.001)	-0.001 (0.001)	-0.078*** (0.022)	-0.075*** (0.023)
Controls	No	Yes	No	Yes
Judge fixed effects	Yes	Yes	Yes	Yes
Observations	8,446	8,446	8,446	8,446

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 A6: Testing effect of regressors on distribution of judge errors

	(1)	(2)	(3)
<i>Coefficients β_1 affecting judge standards γ_1</i>			
End-of-week decision	0.056*** (0.001)	0.072*** (0.003)	0.058*** (0.002)
Hearing scheduled over lunch	0.543*** (0.037)	0.491*** (0.027)	0.520*** (0.021)
<i>Coefficients ψ_1 affecting judge inaccuracy σ_1</i>			
End-of-week decision		0.024*** (0.002)	
Hearing scheduled over lunch			0.072*** (0.008)
SD of γ_1	0.875	0.832	0.873
SD of σ_1	0.695	0.593	0.648

Reports coefficients for choice model $\mathbb{1}[r_i > X_{ijs}\beta_s + \gamma_{js} + \tilde{\varepsilon}_{ijs}(W_{ijs})]$, $\tilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, e^{\sigma_{js} + W_{ijs}\psi})$. All models include controls for time/day of decision in β , 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 A7: Polynomial monotonicity p -values, by group definition and degree

	(1)	(2)	(3)	(4)	(5)
Degree	1	2	3	4	5
<hr/> Group definition <hr/>					
Male-Asia-Montreal	1	.999	.147	.003	0
Male-Africa-Montreal	1	1	.963	0	0
Male-Asia	1	.984	.41	0	.001
Male-Africa	1	1	.769	0	0
Male-Montreal	1	1	.91	.019	0
Male-Asia-Africa	1	.999	.509	0	.001
Asia-Montreal	1	.998	.015	.001	0
Africa-Montreal	1	1	.934	0	0
Male	1	1	.621	0	0
Asia	1	.818	.15	0	.002
Africa	1	1	.820	0	.002
Montreal	1	1	.873	.005	0

Displays p -values for polynomial test of first-stage monotonicity for different group definition, where each group is the intersection of the listed characteristics.

Table A8: Second-round approval on model approval probability and judge-pair FEs, no-regressor model

	Judge-pair round FEs		Judge-pair FEs	
	(1)	(2)	(3)	(4)
Model approval probability	0.918*** (0.157)	0.920*** (0.157)	1.011*** (0.0487)	1.010*** (0.0487)
Model controls	No	Yes	No	Yes
Mean approval	0.44	0.44	0.44	0.44
F-stat for judge pairs	1.01	1.01	0.98	0.98
p -value	0.718	0.746	0.820	0.787
SD of judge-pair EB means	0.000	0.000	0.008	0.007
Observations	8,179	8,179	8,179	8,179

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 A9: First-round judge accuracy by experience and workload, no-regressor model

	(1)	(2)	(3)	(4)
<i>Coefficients ψ_1 affecting judge inaccuracy σ_1</i>				
Experience > 1 year	-0.816*** (0.020)	-1.165*** (0.052)	-0.781*** (0.024)	-0.312 (0.212)
Experience > 5 years	-0.317*** (0.021)	-0.516*** (0.026)	-0.469*** (0.030)	0.035 (0.193)
Experience > 10 years	-0.368*** (0.017)	-0.631*** (0.043)	-0.886*** (0.027)	-0.925*** (0.197)
Log caseload			0.180*** (0.004)	
Log caseload (≤ 5 yrs exp)				0.281*** (0.051)
Log caseload (> 5 yrs exp)				0.036 (0.070)
SD of γ_1	0.692	0.779	0.735	1.023
SD of σ_1	1.769	1.599	1.623	1.417
Second-round experience control	Yes	Yes	Yes	Yes
Career number of cases	No	No	No	No

Reports coefficients for decision model $\mathbb{1}[r_i > X_{ijs}\beta_s + \gamma_{js} + \tilde{\varepsilon}_{ijs}(W_{ijs})]$, $\tilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, e^{\sigma_{js} + W_{ijs}\psi})$. 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 contrast to the main results reported in Table 4, models are estimated without using regressors for identification. Standard errors clustered at the level of the first stage judge. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A10: Inaccuracy before and after judge selection reform, no-regressor model

	Baseline			Experience control in σ_1		
	(1)	(2)	(3)	(4)	(5)	(6)
Appointed after reform (=1)	-0.288 (0.185)	-0.804** (0.387)	-0.887** (0.417)	-1.050*** (0.282)	-1.165*** (0.370)	-1.213*** (0.390)
Liberal appointee (=1)			-0.00267 (0.195)			0.0356 (0.134)
Male judge (=1)			-0.340 (0.209)			-0.154 (0.146)
Year appointed	No	Yes	Yes	No	Yes	Yes
Pre-reform mean	1.23	1.23	1.23	1.59	1.59	1.59
N judges	53	53	53	53	53	53

Estimated with Hanushek (1974) correction for estimated dependent variable. Dependent variable is inaccuracy σ_{j1} , which is estimated from decision model $\mathbb{1}[r_i > \gamma_{js} + X_{ijs}\beta_s + \tilde{\varepsilon}_{ijs}(W_{ijs})]$, $\tilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\tilde{\sigma}_{js} + W_{ijs}\psi_s)$. In contrast to the baseline models reported in Table 5, 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, β_s and ψ_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 A11: Judge standards γ_1 and inaccuracy σ_1 on lawyer ratings, no-regressor model

	γ_1 (mean=1.87, SD=.84)			σ_1 (mean=.59, SD=.47)		
	(1)	(2)	(3)	(4)	(5)	(6)
Lawyer favorability rating, SD	-0.364*		-0.303	-0.083		-0.027
	(0.183)		(0.190)	(0.093)		(0.094)
Lawyer unpredictability rating, SD		0.257**	0.148		0.147**	0.137**
		(0.108)	(0.104)		(0.054)	(0.057)
Observations	73	73	73	73	73	73

Reports linear regressions of model coefficients on survey responses, estimated with Hanushek (1974) correction for estimated dependent variable. Decision model is $\mathbb{1}[r_i > \gamma_{js} + \tilde{\varepsilon}_{ijs}]$, $\tilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\tilde{\sigma}_{js} + W_{ijs}\psi_s)$. In contrast to the baseline model reported in Table 6, 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 β_s and ψ_s . Standard errors clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A12: Lawyer characteristics, survey respondents vs lawyer population

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

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 A13: Model coefficients on survey lawyer responses, controlling for other survey responses

	γ_1 (mean=1.84, SD=.84)			σ_1 (mean=.57, SD=.45)		
	(1)	(2)	(3)	(4)	(5)	(6)
Lawyer favorability rating, SD	-0.359*		-0.294	-0.069		-0.012
	(0.196)		(0.234)	(0.092)		(0.119)
Lawyer unpredictability rating, SD		0.269**	0.163		0.133**	0.130*
		(0.109)	(0.134)		(0.052)	(0.075)
Observations	69	69	69	69	69	69

Reports linear regressions of model coefficients on survey responses, estimated with Hanushek (1974) correction for estimated dependent variable. Decision model is $\mathbb{1}[r_i > \gamma_{js} + \tilde{\varepsilon}_{ijs}]$, $\tilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\tilde{\sigma}_{js} + W_{ijs}\psi_s)$. All models include controls for time/day of decision in β_s , and allow 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 β_s and ψ_s . Standard errors clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.