Judicial Errors: Evidence from Refugee Appeals[*]

Samuel Norris[†]

November 10, 2017

JOB MARKET PAPER

[Most recent version here]

Abstract

Judges with the same overall conviction rate may convict different defendants, which has important implications for the fairness and efficiency of the judicial system. I show how this notion of inconsistency can be identified separately from judicial severity in two-stage court systems by using the second-round judge to validate first-round decisions. Structural estimates of consistency for a sample of Canadian refugee appeal judges are highly correlated with lawyers' surveyed opinions on judge ability. Overall levels of consistency are low. Judges who approve the same share of claimants disagree on the correct decision for 13% of all claimants, and the average judge would have 58% of her approvals rejected by a similarly-severe colleague. However, judges become much more consistent as they gain experience, with the largest gains coming in the first year. Across judges, consistency is higher for judges appointed after a 1988 reform designed to reduce politically-motivated appointments. One ramification of inconsistency is that many claimants who would be successful in the second round are denied in the first round. If all claimants were given a second-round hearing, approximately 11,400 cases would be successful over 1995-2012, versus 3,700 under current policy.

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

[†] Department of Economics, Northwestern University. norris@u.northwestern.edu

1 Introduction

The justice system is a major institution in all developed countries. In the US alone, there are approximately 7 million felons and ex-felons under court supervision (Glaze and Parks, 2011), and 47 million new non-traffic cases filed in state courts each year (Bureau of Justice Statistics, 2006). Other quasi-judicial institutions, such as the system of Social Security Disability Insurance examiners who decide SSDI eligibility, routinely make decisions worth tens of thousands of dollars (Maestas et al., 2012).

The efficiency and fairness of the courts has far-reaching consequences. Coase (1960) makes the general case that unclear or ambiguous property rights often lead to inefficient economic outcomes, and Porta et al. (1998) the more specific point that inefficient courts increase transaction costs and reduce aggregate investment. However, evidence on the overall efficacy of the courts remains limited. In specific situations, there is compelling evidence that judicial decisions are affected by non-relevant factors like upcoming elections (Canes-Wrone, Clark, and Kelly, 2014), inter-communal violence unrelated to the crime (Shayo and Zussman, 2010), the previous decision (Chen, Moskowitz, and Shue, 2016), the timing of the hearing relative to lunch (Danziger, Levav, and Avnaim-Pesso, 2011) and the winner of last night's football game (Eren and Mocan, 2016). All of these features of the justice system contribute to randomness in the decision. One way to measure overall inefficiency would be to add up all these examples and any others a researcher could measure. In this paper, I take a different approach and directly measure aggregate randomness in judicial decision-making as well as the reliability of individual judges. Do judges differ only in *leniency*, the share of defendants they incarcerate? Or are judges also *inconsistent* — would judges with the same incarceration rate incarcerate the same defendants? Do judges vary in consistency? The tools I develop in this paper allow me to separately identify judge-level leniency and consistency, which facilitate evaluations of the efficiency of the judicial system (are guilty defendants likely to be incarcerated?) and allow the researcher to measure the success of reforms. Consistency is closely related to the montonicity assumption of examiner-assignment IV designs, and my results shed light on potential biases in this increasingly-common identification strategy (Dahl, Kostol, and Mogstad, 2013; Mueller-Smith, 2014).

My conception of consistency is a simple generalization of the usual index model of judicial decision-making, where judges perfectly observe the strength of each claimants case and approve them if it is larger than some judge-specific threshold.[1] In my model, judges observe case quality with error; the size of the distribution of this error is inconsistency. Judicial behavior is thus summarized by a judge-specific threshold and a judge-specific error distribution. In many environments this is not identifiable: if we observe only judge-specific approval rates, a judge who perfectly selected all

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

claimants meeting the legal standard would be indistinguishable from a judge who approved the same share of claimants but flipped coins to do so.

Identification relies on two distinct institutional characteristics. The first is random assignment of judges to cases, which is common in many court systems and ensures that underlying case characteristics are uncorrelated with judge characteristics. The second is that the decision is made by two judges acting independently but using similar criteria. This latter condition can be met by identical standards for each judge, but in my setting and in many others is satisfied by the requirement that claimants be recognized as having an arguable case by one judge (in legal parlance, granted leave) before being given a full hearing in front of another. I use the second-stage decision to check the accuracy of the first-stage decision — if there are two first-round judges who approve the same number of first-round claimants, the more consistent judge will have a higher share of her claimants approved by the second-round judge. Similarly, approval rates for consistent second-round judges — who can easily distinguish between high and low quality cases — increase more than for inconsistent judges when the severity of the first-round judge (and corresponding case quality of the approved) increases. I show that the rest of the model, including the distribution of unobserved case strength, can also be identified using regressors that affect judge leniency. I build a structural model combining the two sources of identification that identifies leniency and consistency for each judge, as well as the distribution of underlying case quality. The model is nonparametrically identified, and can be tractably estimated via maximum likelihood under parametric restrictions.

This paper is related to two different literatures. First, the idea behind my identification strategy — that observing multiple decision-makers on the same case is informative about the accuracy of decisions — appears in many different contexts. In a reduced-form sense, Frakes and Wasserman (2014) use patent decisions from non-US patent offices to generate an independent measure of patent quality, then examine how the quality of granted patents for US examiners changes as they are given less time to make a decision. Another set of papers grapples with selection-type models where outcomes are observed only conditional on treatment or some other agent decision. Chandra and Staiger (2011) develop a model where hospitals both vary in their ability to treat heart attack patients, and choose which ones to treat. They identify hospital-level treatment effects using a structural model that interprets patient survival measures through a lens of treatment selection. Abaluck et al. (2016) study how doctors choose which patients to send for imaging tests for pulmonary embolism. Since the test reveals whether the patient actually has the disease, high test yield rates (conditional on share of patients sent for a test) are an indication of good allocation of tests across patients. Similarly, Anwar and Fang (2006) develop a hit rate test that compares the proportion of black and white drivers who are found to be transporting drugs after a vehicle search to test for racial bias in the search decision among police officers. Closer to my context, Alesina and Ferrara (2011) use appeals in capital sentencing to test for racial bias under the assumption that higher courts are less racially biased than lower ones. I expand on this work by showing how the consistency of both the first- and second- round decision-maker can be identified, even when the researcher does not have access to objective measures of the truth. The model is applicable to any situation where two potentially fallible decision-makers are independently making a similar decision. It can be used to understand which decision-makers are most consistent, and what factors increase consistency.

Second, I contribute to the literature on judicial decision-making. Previous research has documented large variation in conviction rates across judges under random assignment of cases (Aizer and Doyle, 2013; Bhuller et al., 2016; Rehaag, 2007), which implies that there are cases where judges would disagree on the correct decision. Fischman (2013) shows that the share of cases a pair of judges would disagree on can be bounded using Fréchet inequalities when the researcher does not observe the two judges making decisions on the same cases.[2] Another strand of research looks at multi-judge panels, although strategic interactions and consensus norms among judges make modelling much more difficult than in my context (Epstein, Landes, and Posner, 2013; Fischman, 2008). Finally, a directly relevant paper is Partridge and Eldridge (1974), who provide 50 district court judges with identical cases and compare the judges' hypothetical sentences. Although there is some fear the study lacks external validity because the cases were hypothetical, the results are interesting and anticipate my own findings. They show that there is a high degree of disparity in the sentences given for the same case, and that this disparity is not primarily caused by individual judges being consistently lenient or consistently harsh. As they put it, "if there are indeed hanging judges and lenient ones — and it would appear that there are a few — their contribution to the disparity problem is minor compared to the contribution made by judges who cannot be so characterized." This suggests that judicial inconsistency may be large in the sense of my model.

I apply my model to judicial review of refugee cases at the Federal Court (FC) of Canada. The FC is the only point of appeal for claimants who have been rejected for refugee status by administrative decision-makers at the Immigration and Refugee Board (IRB), and is seen as a crucial backstop that ensures the fairness of the overall refugee system.[3] The stakes are high. As noted in Rehaag (2012), "if errors in first-instance refugee determinations at the [IRB] are not caught and corrected through judicial review, refugees may be deported to countries where they face persecution, torture or death." The judges are experts in dealing with refugee cases; about 70% of their caseload is refugee appeals. Nonetheless, I find low levels of consistency between judges, corresponding to a meaningful impact on decisions and outcomes. On average, judges who approve the same share of claimants disagree on the correct decision for 13.2% of cases, or an astonishing 57.6% of the cases

[2] In contrast, I conceptualize inconsistency as two judges with the same approval rate making different decisions, which is a necessarily more conservative definition.

[3] In legal terminology, judicial review has a different meaning than the more-familiar 'appeal;' it refers specifically to the judicial oversight of an administrative decision. For ease of language I will use the term appeal rather than judicial review throughout this paper.

they approve. Suggestive evidence indicates that this is due to idiosyncratic observational errors, rather than different judging ideologies or statutory interpretations.

The lack of consistency can also be understood as a failure of first-round judges to pick the claimants that will be successful at a full hearing in the second round. If all claimants were given a full hearing, my results suggest that 19.4% of IRB denials of refugee status would be overturned, rather than the 6% that is ultimately successful under the current system.[4] This difference amounts to approximately 7,700 families over my study period.

I survey refugee lawyers about judge quality, and validate the model by showing that survey responses are correlated with my measures of consistency and leniency. Consistency improves dramatically during the first year of experience, and continues to improve at a slower rate for at least ten years. Judges during the first five years of experience are less consistent during periods of high workload, but experienced judges are unaffected by workload.

In 1988 a law was passed to make it more difficult for the government to appoint unqualified judges. The reform, which gave a committee of legal experts veto power over candidates, had the intended effect of reducing the number of newly-appointed judges with ties to the party in power (Hausegger et al., 2010; Russell and Ziegel, 1991). I find that it dramatically improved judge consistency, implying that reforms to judicial selection processes can have meaningful effects on judicial outcomes and efficiency.

In a final section, I show how my model can be used to construct counterfactual judge assignment regimes that minimize workload while approving the same number of claimants and maintaining the case quality of the approved. I find that the Federal Court could reduce refugee workload by approximately 18% while approving similarly-qualified claimants, saving at least $4.4 million in judge salaries alone over my study period.[5]

The paper proceeds in five parts. In Section 2 I present the model and discuss identification. Section 3 contains a discussion of the institutional background and data, and Section 4 the results. Section 5 concludes.

2 Model and identification

I discuss the institutional setting in detail in Section 3. To fix ideas before presenting the model, the Federal Court hears appeals from claimants who have been denied refugee status by the government. Enormous variation in approval rates for government decision-makers suggests that there is a long

[4] Theoretical work predicts this finding to some degree. Sah and Stiglitz (1986) compare a centralized decision-making process comparable to the Federal Court's with more decentralized processes and show that centralization leads to greater sensitivity to low-quality decision makers.

[5] Alternatively, judge assignments could be reshuffled to maximize the number of successful claims while holding workloads constant and maintaining the case quality of the approved. The problem is approximately symmetric, so the same judicial resources could be used to increase the number of approvals by 19%.

tail of claimants who would have been approved had they been assigned an alternative decision-maker, and should be successful on appeal. Decisions at the Federal Court are made in a two-stage process, where the criteria for a first-round decision is whether a claimant would have an arguable case in the second round. In this sense, the criteria are the similar in both rounds. The second round proceeds only if there is a first-round approval, and includes a hearing where lawyers for both sides argue about the case but do not introduce new evidence. Judges are quasi-randomly assigned in both rounds.

2.1 Model

The court receives a flow of applicants for refugee status. Strength of case for each applicant i can be represented by a scalar, $r_i \sim F_r$. To be approved as a refugee, a claimant must be approved by two consecutive judges. If she is denied by the first judge, her case is not seen by the second judge. Formally, in stages $s = 1, 2$, judges $j = 1...J$ approve the claimant if

$$r_i > \varepsilon_{ijs}(X_{ijs}, W_{ijs}) = \gamma_{js} + X_{ijs}\beta_s + \widetilde{\varepsilon}_{ijs}(W_{ijs}) \tag{1}$$

where $\widetilde{\varepsilon}_{ijs} \sim G_{js}$ and $\exists x_{ijs} \in X_{ijs}$ s.t. $x_{ijs} \notin W_{ijs}$. Judge *leniency* is captured by γ_{js}; high levels of γ_{js} mean that fewer claimants are approved. This threshold can be adjusted by X_{ijs}.[6]

Judge *consistency* is defined by the distribution of $\widetilde{\varepsilon}_{js}(W_{ijs})$. For perfectly consistent judges, $\widetilde{\varepsilon}_{js}(W_{ijs}) = 0$. Then, the decision problem is non-stochastic for a given value of r_i: $P[r_i > \varepsilon_{js}(X_{ijs}, W_{ijs})] = P[U > F_r(\gamma_{js} + X_{ijs}\beta_s)] = 1 - F_r(\gamma_{js} + X_{ijs}\beta_s)$, and so any two judges with the same overall approval rate would either both approve or both reject any claimant with given quality $\tilde{r}_i$ (this is the standard model of judicial decision-making). Judges become less consistent as the distribution of $\widetilde{\varepsilon}_{js}(W_{ijs})$ widens. Across judges, more consistent judges are more likely to approve claimants with a strong case (high r_i). I operationalize consistency by comparing judges with the same approval rate. In the first round, the approval rate is

$$P[r_i > \varepsilon_{j1}(X_{ijs}, W_{ijs})] = \int G_{j1}(r_i - \gamma_{j1} - X_{ij1}\beta_1) f_r dr \tag{2}$$

Judge A is *comparable* to judge B when $P[r_i > \varepsilon_{A1}] = P[r_i > \varepsilon_{B1}]$. Then, he is more consistent than judge B if there exists a point of single-crossing v such that:

1. $G_{A1}(v - \gamma_{A1}) = G_{B1}(v - \gamma_{B1})$

2. $\forall\ w > v, G_{A1}(w - \gamma_{A1}) \geq G_{B1}(v - \gamma_{B1})$, with a strict equality for some w with $f_r(w) \neq 0$

[6] It is mathematically equivalent if X_{ijs} shifts the distribution of r_i without otherwise affecting the distribution, but conceptually difficult to imagine what X_{ijs} might do this.

3. $\forall\ w < v, G_{A1}(w - \gamma_{A1}) \leq G_{B1}(v - \gamma_{B1})$, with a strict equality for some w with $f_r(w) \neq 0$

In words, this definition is straightforward: a consistent judge is more likely to approve high-quality claimants, and less likely to approve low-quality claimants. I assume that for any pair of comparable judges, one is more consistent than the other (this can be thought of as a single-crossing property for the error CDFs G_{js}). My definition is similar to the concept of screening in Sah and Stiglitz (1986), where they define A as more *discriminating* than B when $\partial G_{A1}(v - \gamma_{A1})/\partial v > \partial G_{B1}(v - \gamma_{B1})/\partial v$. In my model, the more consistent judge is (weakly) more discriminating at the point of single-crossing.

The joint probability of approval in the first and second rounds (where in the second round the claimant faces a potentially different judge k) is

$$P[r_i > \varepsilon_{j1}(X_{ij1}) \cap r_i > \varepsilon_{k2}(X_{ik2})] = \int G_{j1}(r_i - X_{ij1}\beta_1 - \gamma_{j1}) G_{k2}(r_i - X_{ik2}\beta_2 - \gamma_{k2}) f_r dr \tag{3}$$

2.2 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 of varying severity, and regressors that shift judge thresholds. I consider each in turn.

2.2.1 Judge-assignment identification

Take two judges with the same first-round approval rate, A and B. Then, a higher share of the more consistent judge's claimants will be ultimately approved by a common second-round judge, C. Suppose that judge A is more consistent. Then, abstracting away from covariates X_{ij1} and substituting $\widetilde{G}_{js}(r) = G_{js}(r - \gamma)$ for clarity, this can be seen by noting that:

$$\begin{aligned}
&\left(P[r > \varepsilon_{C2} | r > \varepsilon_{A1}] - P[r > \varepsilon_{C2} | r > \varepsilon_{B1}]\right) P[r > \varepsilon_{B1}] \qquad (4)\\
=&P[r > \varepsilon_{A1} \cap r > \varepsilon_{C2}] - P[r > \varepsilon_{B1} \cap r > \varepsilon_{C2}]\\
=&\int \left[\widetilde{G}_{A1}(r) - \widetilde{G}_{B1}(r)\right] \widetilde{G}_{C2}(r) f_r dr\\
=&\int_{-\infty}^{z} \left[\widetilde{G}_{A1}(r) - \widetilde{G}_{B1}(r)\right] \widetilde{G}_{C2}(r) f_r dr + \int_{z}^{\infty} \left[\widetilde{G}_{A1}(r) - \widetilde{G}_{B1}(r)\right] \widetilde{G}_{C2}(r) f_r dr\\
>&\int_{-\infty}^{z} \left[\widetilde{G}_{A1}(r) - \widetilde{G}_{B1}(r)\right] \widetilde{G}_{C2}(z) f_r dr + \int_{z}^{\infty} \left[\widetilde{G}_{A1}(r) - \widetilde{G}_{B1}(r)\right] \widetilde{G}_{C2}(z) f_r dr
\end{aligned}$$

$$= \widetilde{G}_{C2}(z) \int \left[\widetilde{G}_{A1}(r) - \widetilde{G}_{B1}(r)\right] f_r dr = \ 0$$

where z is the point of single-crossing of $\widetilde{G}_{A1}$ and $\widetilde{G}_{B1}$. 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 $\widetilde{G}_{C2}(\cdot)$; since the second-round judge is more likely to approve high-r claimants than low-r claimants, his decisions are informative about which first-round judge has chosen higher-quality first-round claimants. The last equality comes from the comparability of judges A and B, underlining that this is a local result: it tells us which judges are more consistent, but compares only judges with similar approval rates.

Identification of second-round consistency follows a slightly different route, because we do not have a third round to use as a check. Instead, I attain comparability from a *non-limiting* judge. Suppose we are trying to determine which second-round judge, A or B, is more consistent. I assume that there is a known first-round judge D that approves nearly anyone, and a known first-round comparison judge C. Formally, I require that $\widetilde{G}_{C1}(\cdot)/\widetilde{G}_{D1}(\cdot)$ is monotonically increasing wherever $\widetilde{G}_{A2}(\cdot) \neq \widetilde{G}_{B2}(\cdot)$. This is trivially satisfied when $G_{D1}(\cdot) = 1$ (judge D literally approves everyone), and can be satisfied when judge D is fairly consistent and has a very low threshold γ relative to judge C.

Define judge A and B as second-round comparable if they have the same second-round approval rate conditional on first-round approval by judge D; $\int \widetilde{G}_{D1}(r)\widetilde{G}_{A2}(r) f_r dr = \int \widetilde{G}_{D1}(r)\widetilde{G}_{B2}(r) f_r dr$. Then, if judge A's second-round approval rate increases more than judge B's for decisions conditional on judge C's first-round approval (vs. judge D's), judge A is more consistent than judge B. This can be seen by the following derivation,

$$\begin{aligned}
&\left(P[r > \varepsilon_{A2} | r > \varepsilon_{C1}] - P[r > \varepsilon_{B2} | r > \varepsilon_{C1}]\right) P[r > \varepsilon_{C1}] \qquad (5)\\
=&P[r > \varepsilon_{C1} \cap r > \varepsilon_{A2}] - P[r > \varepsilon_{C1} \cap r > \varepsilon_{B2}]\\
=&\int \widetilde{G}_{C1}(r) \left[\widetilde{G}_{A2}(r) - \widetilde{G}_{B2}(r)\right] f_r dr\\
=&\int_{-\infty}^{z} \frac{\widetilde{G}_{C1}(r)}{\widetilde{G}_{D1}(r)} \widetilde{G}_{D1}(r) \left[\widetilde{G}_{A2}(r) - \widetilde{G}_{B2}(r)\right] f_r dr + \int_{z}^{\infty} \frac{\widetilde{G}_{C1}(r)}{\widetilde{G}_{D1}(r)} \widetilde{G}_{D1}(r) \left[\widetilde{G}_{A2}(r) - \widetilde{G}_{B2}(r)\right] f_r dr\\
>&\int_{-\infty}^{z} \frac{\widetilde{G}_{C1}(z)}{\widetilde{G}_{D1}(z)} \widetilde{G}_{D1}(r) \left[\widetilde{G}_{A2}(r) - \widetilde{G}_{B2}(r)\right] f_r dr + \int_{z}^{\infty} \frac{\widetilde{G}_{C1}(z)}{\widetilde{G}_{D1}(z)} \widetilde{G}_{D1}(r) \left[\widetilde{G}_{A2}(r) - \widetilde{G}_{B2}(r)\right] f_r dr\\
=&0
\end{aligned}$$

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

2.2.2 Regressors and identification

The between-judge comparisons that I discuss in the previous section are local; they measure relative consistency for judges with similar approval rates. To compare judges who approve different shares of claimants and to identify the scale of judge errors $\widetilde{\varepsilon}_{ijs}$ without resorting to functional form assumptions, additional large-support continuous regressors are required. These regressors affect judge thresholds 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 (ie, $\widetilde{\varepsilon}_{ijs} - r_i$) for each round. I then assume that at least one component of β_s is the same between rounds, tying down the relative size of the composite errors and identifying the distribution of r_i separately from $\widetilde{\varepsilon}_{ijs}$. In a parametric model, instruments are not strictly necessary (the model is mechanically identified), but provide a source of identification beyond functional form and judge randomization. A full proof of identification is in Chen et al. (2000); which I reframe in terms of my model in Appendix Section A1.

2.3 Interpretation

2.3.1 Ideological versus observational errors

The innovation of this model is to separately identify judge thresholds γ_{js}, the distribution of unobserved case strength r_i and the case-judge-stage error $\widetilde{\varepsilon}_{ijs}$. The model guarantees that individuals with the same scalar quality factor r_i have the same overall approval probability, and that judge errors $\widetilde{\varepsilon}_{ijs}$ are uncorrelated with both case strength r_i and $\widetilde{\varepsilon}_{ijs'}$ for the other stage s'. A useful way to think of r_i is as a measure of average quality, where the average is taken across judges.

In the first round, $\widetilde{\varepsilon}_{ij1}$ can be decomposed into two conceptually distinct components: permanent differences in how a judge interprets the law relative to other judges, and pure observational errors. Formally,

$$\widetilde{\varepsilon}_{ij1} = u_{ij} + e_{ij1} \tag{6}$$

The first component, u_{ij}, represents permanent disagreements, or inter-rater reliability. As I discuss in Section 3, refugee appeals at the Federal Court are assessed along both procedural and substantive lines. In other words, one judge might always reject a claimant who has a strong procedural case and a weak substantive one, while a different judge who weighs substantive considerations more heavily might always approve him. This term also includes differential bias along racial or gender lines. u_{ij} is therefore a measure of how a judge's weighting of different facets of a case differs from the consensus.

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

A natural question concerns the relative size of r_i, u_{ij} and e_{ij1}. My model identifies the relative variance of r_i versus the composite error $u_{ij} + e_{ij1}$, but does not directly estimate the size of u_{ij} versus e_{ij1}. However, the strength of the *additional* predictive power of judge identity has strong implications for the relative size of the errors. Suppose that the composite error was mostly inter-rater differences. Then, one would expect that some pairs of judges would both value the same type of cases (for example, cases that were particularly strong on the procedural merits, or involved an Asian claimant). By definition, the probability of approval in the second round for a claimant with quality r_i and judges j and k (and suppressing extra regressors) is

$$P_{jk} = P[\text{Approval by } j | \text{Approval by } k \text{ in } 1^{\text{st}}] = \frac{P[r_i > \gamma_{j2} + u_{ij} + e_{ij2} \cap r_i > \gamma_{k1} + u_{ik} + e_{ik1}]}{P[r_i > \gamma_{k1} + u_{ik} + e_{ik1}]}$$

If two judges have a similar judging ideology, then they will both be predisposed to treat the same case either more or less positively than would be predicted by the factor refugee quality r_i, their γ's and their σ's. More formally, index P_{jk} by the correlation in ideological errors u_{ij} and u_{ik}, ρ_{jk} (the model as estimated implicitly assumes $\rho = 0$). It is simple to show that $P_{jk}(\rho_{jk})$ is increasing in ρ_{jk}. A reduced-form test for whether there are important judge-pair agreements in ideology is to regress

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

where $P_{jk}(0)$ is calculated from the model.[7] Under the null of no correlations in errors between judge pairs, the judge-pair fixed effects ν_{jk} should be jointly insignificant. This is a joint test of *all* the reasons pairs of judges could disproportionately agree or disagree — importantly, it also includes that some pairs of justices may disproportionately trust each others judgment — but failure to reject suggests that inter-rater differences are not large. This is turn suggests that test-retest errors $\breve{\varepsilon}_{ijs}$ are larger than inter-rater differences $\bar{\varepsilon}_{ij}$. I implement this test in Section 4.6, and fail to reject the null of no judge-pair effects.

[7] Note that the model assumes $\widetilde{\varepsilon}_{ij1} \perp \widetilde{\varepsilon}_{ik2}$. This assumption might at first glance be odds with allowing judge-pair correlations. The important distinction is that the model assumes $\widetilde{\varepsilon}_{ij1} \perp \widetilde{\varepsilon}_{ik2}$ without conditioning on the identity of the judges — it imposes that the judge errors are uncorrelated conditional on the index threshold.

2.3.2 1$^{\text{st}}$ versus 2$^{\text{nd}}$ round errors

The decomposition of errors into inter-rater and test-retest components is complicated in the second stage by the possibility that the judges gain additional information about the case in the second-round hearing. As I discuss in Section 3.2, there is a full hearing in the second round (in the first they just review documents), and judges may learn things about the case that color their views of its strength. This is conceptually distinct from both observational errors e_{ij2} and inter-rater disagreements u_{ij} in that this new information could reflect information about the true merits of the case. In the second round the error can be decomposed

$$\widetilde{\varepsilon}_{ij2} = u_{ij} + e_{ij2} + \mathcal{I}_{ij2} \tag{8}$$

where $\mathcal{I}_{ij2}$ is an information shock. This shock should be thought of as realized 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 interpretation of the other components is analogous to the first round: u_{ij} is a judge-level bias term that reflects judge-level tendencies to accept certain arguments or types of cases, and e_{ij2} is observational error.

The potential presence of the information shock $\mathcal{I}_{ij2}$ complicates interpretations of the size of the distribution of $\widetilde{\varepsilon}_{ij1}$, because not all of the variation is attributable to judge errors — some may reflect new information. A wide distribution of $\widetilde{\varepsilon}_{ij2}$ could in fact reflect a particularly perceptive judge. For this reason, I focus most of my discussion on $\widetilde{\varepsilon}_{ij1}$.

2.3.3 Interpreting the magnitude of inconsistency

The most straightforward reduced form measure of judicial inconsistency is the share of claimants that judges disagree on. This comparison is particularly sharp between judges who approve the same share of claimants, because perfect consistency for both judges in this situation means they would not disagree on any cases. Suppressing covariates so that $\varepsilon_{ijs} = \gamma_{js} + \widetilde{\varepsilon}_{ijs}$, for judges j and k in the first round, this can be calculated as

$$\int\int\int \Big\{ \mathbb{1}[r_i > \varepsilon_{ij1}]\mathbb{1}[r_i < \varepsilon_{ik1}] + \mathbb{1}[r_i < \varepsilon_{ij1}]\mathbb{1}[r_i > \varepsilon_{ik1}] \Big\} f_r \; d\widetilde{\varepsilon}_{ij1} \; d\widetilde{\varepsilon}_{ik1} \; dr \tag{9}$$

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

As I will show in Section 4.6, inconsistency seems to be driven mostly by observational errors rather than ideology. I interpret this to mean that $\widetilde{\varepsilon}_{ij1}$ is likely to be uncorrelated across judges in the same round. In that case, Equation 9 can be used to calculate the disagreement rate under the assumption that $\widetilde{\varepsilon}_{ij1} \perp \widetilde{\varepsilon}_{ik1}$. I refer to this as *uncorrelated disagreement.*

Alternatively, I bound the size of the disagreement and find the minimum level of disagreement for any joint distribution of $\widetilde{\varepsilon}_{ij1}$ and $\widetilde{\varepsilon}_{ik1}$. For each r_i, a pair of judges disagrees with at least probability $|G_{j1}(r_i - \gamma_{j1}) - G_{k1}(r_i - \gamma_{k1})|$. Total disagreement can then be bounded by integrating over the distribution of case strength r_i. By construction, *bounded disagreement* is a conservative measure of disagreement, but will be larger than zero whenever judges vary in the distribution of their observational error.

3 Institutional Background and Data

This section describes the refugee adjudication system as it existed during the study period. Initial refugee decisions in Canada are made by an independent administrative body known as the Immigration and Refugee Board (IRB). The IRB is not itself amenable to analysis because the data is mostly unavailable and the procedures to assign adjudicators to cases are opaque (and non-random). My entire analysis therefore concerns the Federal Court, which hears appeals of IRB decisions and fits the institutional criteria necessary for identification. However, I begin by describing the IRB in enough detail to contextualize the distribution of initially denied claimants who appeal to the Federal Court. I describe the Federal Court and the procedure the government uses to select justices for the Court, then introduce the data and discuss estimation.

3.1 Immigration and Refugee Board

Initial screening of inland refugee claims is conducted by the Members of the IRB, who are tasked with evaluating whether the claimant meets the statutory definition of a refugee: “a person who, by reason of a well-founded fear of persecution for reasons of race, religion, nationality, membership in a particular social group or political opinion, is outside each of their countries of nationality and is unable or, by reason of fear, unwilling to avail themselves of the protection of each of those countries.” Claims are non-randomly assigned to Members with expertise relevant to the type of case; this expertise is usually in terms of either the country of origin of the claimant or the stated reason for the claim. The Members are political appointees rather than long-term, professional bureaucrats.

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

Claimants who have been rejected for refugee status may apply to the Federal Court for judicial review. Approximately 65% of denied claimants file an appeal, which allows most claimants to stay in Canada until the Federal Court makes its final decision.[8]

IRB procedures for making refugee determinations were broadly consistent from 1995 until December 15, 2012, when an administrative appeal division partially supplanted the review work of the Federal Court (Grant and Rehaag, 2015). The only major policy change in this period concerned the composition of the IRB panel that made the decision. For refugee claims submitted before June 28, 2002, standard procedure was for the case to be heard by a two-Member panel. If either member recommended approval, refugee status would be granted. Upon consent of the claimant, the case could be heard by a single Member, and by 2002 this practice was common (Dauvergne, 2003). However, the claimant often knew which Member would be making the decision if they agreed to a single-Member panel. Ostensibly they would be less likely to let the decision on their refugee status be made by a Member with a low approval rate, meaning that they had some ability to pick who would decide their refugee status. After the implementation of the Immigration and Refugee Protection Act (IRPA) in 2002, all cases were heard by a single Member. This is important because it suggests that the distribution of case strenth for the rejected claimants who appeal to the Federal Court changed after IRPA came into affect; more high-quality claimants may have been rejected, skewing the distribution further to the right. To allow for this possibility, I allow for the distribution of case strength r_i to vary before and after IRPA. More details are in Section 3.5.

3.2 Federal Court responsibilities and protocol

The Federal Court is a national court with jurisdiction over certain issues related to the federal government. The 33 judges of the court hear cases related to intellectual property, maritime law, and aboriginal law, but about 70% of their caseload is devoted to appeals of IRB decisions.

The first round of the process is the leave stage, where a single quasi-randomly assigned judge is tasked with deciding whether a claimant has an "arguable case" to make in a full second-round hearing. In my model, the distribution of case quality r_i is identified because it is (imperfectly)

[8] The IRB occasionally rules a refugee application was "without merit." In that case, removal can occur before judicial review at the FC.

observed by both judges. The arguable-case standard is therefore important because it maps the ultimate standard from the second stage into the first stage.

The first-round judge makes her decision after reviewing written records from the IRB decision and briefs written by the lawyers for the claimant (arguing for a second-round hearing) and the government (arguing against). If they decide against judicial review, the claim is rejected and the claimant is usually deported.[9] If the petition for leave is approved, the case goes to a full judicial review (JR) hearing. The judge for the second-round hearing is also quasi-randomly assigned, so usually the second-round judge is someone different. Regardless of the first-stage outcome, the first-round judge does not provide a written explanation for her decision. This may be one reason to expect that second-round decisions will frequently be inconsistent with the first-round approval. It could also contribute to inter-rater inconsistency for first-round judges, since the dearth of written precedent makes it difficult for judges to learn about how their colleagues have ruled on similar cases. It could also contribute to a wider spread in standards (in my model, thresholds γ_j).

The full hearing corresponds to the second stage of my model. During the hearing, the justice questions the lawyers about the contents of their submissions and the IRB records, but very rarely reviews new evidence or calls witnesses. The name of the first-round judge is not immediately available. It is not difficult for the second-round judge to access this information if he wants, but my conversations with judges indicate that they rarely do. To reflect this, I model the second-round decision maker as explicitly ignoring the identity of the first-round judge — the first-round judge affects the second-round decision only through her choice of which claimants to approve, not as a signal to the second-round judge.

In both rounds, judges are not tasked with determining whether the IRB Member made the right decision. Under Canadian law, judges must show deference to administrative decisions. This means that instead of determining whether the "correct" ruling was made, the judge must simply decide whether the government's initial decision was "reasonable" (Rehaag, 2012).[10]

The Federal Court reviews IRB decisions on both substantive and procedural grounds, although the reasonableness standard means that the bar for overturning the decision is high. A substantive ground on which a judge might reverse an IRB decision would be if the Member had ignored credible evidence that a claimant had been tortured. In contrast, procedural reasonableness requires that

[9] There are two legal options for claimants who have been denied leave but do not want to accept the decision, though neither is very common. Beginning the process for either does not forestall removal from Canada. For more details, see Rehaag (2012).

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

the Member collect adequate testimony from the claimant. A judge would be expected to rule in favor of the claimant if there were large procedural violations, even when they believe that the claimant does not actually qualify for refugee status. However, the precise extent to which judges are supposed to weigh substantive and procedural factors is unclear, and it is natural to expect that different judges would differentially consider different aspects of the case. The extent to which this is true is one of the main factors that determines the size of inter-rater inconsistency.

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

Judge assignment works similarly in both stages. For the first stage, judges are assigned to cases using a pre-set schedule; in each office the judges rotate through "leave duty." When enough cases have accrued the court gives the leave duty judge all the outstanding files (usually on Monday), and they are responsible for disposing of all of them. There is no review of the cases before they are given to the judge, and the leave duty schedule is not public. Previous research claims that this assignment is as good as random (Rehaag, 2007); in Section 4 I show that judge leniency is uncorrelated with case or claimant characteristics predictive of success. In the second stage the assignment process is similar; cases are divided between judges who are available for refugee work without review of the contents. A computer program slots hearings into the available times in the judges schedules. Occasionally the same judge will be assigned to a case for both stages by chance. This is potentially important because it implies that between-round errors may be correlated when the same judge is making the decision. Interestingly, this could arise either from inter-rater differences (if a particular claimant has a strong substantive case but a weak procedural one, she would be more likely to succeed in both rounds if she was assigned a judge who heavily weighted substantive aspects) or test-retest errors (a judge might remember the claimant from making the decision in the first round, and so could make the same observational error). I explicitly allow for this by estimating an additional parameter for the correlation between judge errors $\widetilde{\varepsilon}_{ijs}$ in the two rounds whenever the same judge is making the decision in both rounds; more discussion is in Section 3.5.

3.3 Reform to selection of Federal Court justices

Federal Court justices are appointed by the Minister of Justice. Appointments are until the mandatory retirement age of 75.[11] For most of Canadian history the Minister has had nearly unfettered discretion over appointments and has used this power to reward "active supporters of the party in power" (McKelvey, 1985). The only check on the government was a committee of the Canadian Bar Association that offered non-binding advice on the suitability of candidates.

[11] Judges can be removed for misconduct. Occasionally a judge continues to work past the age 75 by having the court classify him as a supernumerary justice. This reclassification has no effect on the work he does.

A major reform in 1988 reduced the discretion of the government in making appointments. The reform created province-level judical advisory councils (JACs) to pre-screen applicants before go went to the Minister for possible selection. The committees were made up of one member of the provincial Law Society, one member of the provincial branch of the bar association, one representative of the provincial chief justice, one representative of the provincial attorney general, and three representatives of the Minister of Justice. The JACs rated each candidate as "highly recommended," "recommended," or "not recommended," and the government could pick judges only from the pool of recommended and highly recommended candidates. The standards concorded well with a lay understanding of what makes a good judge: "'professional competence and experience' (such as proficiency in the law, awareness of racial and gender issues); 'personal characteristics' (ethical standards, fairness, tolerance); and 'potential impediments to appointment' (drug or alcohol dependency, health, financial difficulties)" (Hausegger et al., 2010). Crucially, the direct representatives of the Minister were a minority on the committee, making it difficult to push through wholly unqualified candidates.[12] The standards had bite; only about 40% of candidates were recommended or highly recommended. Although the government could ask a JAC to reconsider a candidate's rating, the reform seems to have reduced the level of patronage. Russell and Ziegel (1991) report that before 1988 at least 47% of appointed judges had some involvement with the ruling Conservative party.[13] Their data comes from reports by surveyed respondents in the legal progression, and so estimates are likely biased down. Though data on post-reform connections to the ruling party are not exactly comparable, Hausegger et al. (2010) search through administrative records and find that after the reform only 30% of newly-appointed judges had donated to the party in power in the five years before their appointment. This is consistent with the new system reducing the number of unqualified party supporters being appointed to the bench, and suggests that the overall consistency of the courts may have improved as a result. I will test this hypothesis in Section 4.9.

3.4 Data

My main data come from Federal Court case reports available on their website.[14] I parsed the data and verified it against a smaller subset professionally transcribed by Rehaag (2012). I use all cases since 1995 that were filed at the IRB before the implementation of the Immigration and Refugee Protection Act (IRPA) on June 28, 2012 (as discussed in Section 3.1, IRPA created a Refugee Appeal Division within the IRB, substantially changing the number and type of refugee appeals

[12] They often share the same name, but provincial political parties in Canada are legally, operationally and usually ideologically independent from the national parties, making coordination on judicial appointments difficult.

[13] The authors distinguish between minor and major involvement. Minor involvement included "minor constituency work, financial contributions, and close personal or professional associations with party leaders;" major involvement running for office, serving as a party official, or active participation in campaigns.

[14] http://cas-cdc-www02.cas-satj.gc.ca/IndexingQueries/infp_queries_e.php

at the Federal Court). I also require that the appeal at the Federal Court was filed before the end of 2012 to ensure that there was enough time for all cases to be disposed of. The dataset has information on the date the case was filed, the Federal Court office that received the application, the name of the leave and judicial review judge, and the ultimate outcome. The office is an important covariate because it strongly predicts outcomes at the court. This is partially because office is correlated with country of origin for claimants, but more because provinces differ in the level of free legal aid provided to claimants. I exclude offices with fewer than 200 cases, leaving Calgary, Montreal, Ottawa, Toronto, and Vancouver.

Using the first name of the claimant, I infer gender using British Columbia and Social Security Administration birth records that contain both first name and gender. To collect information on the country of origin of the claimant, I link the court records to the subset of available IRB case files.[15] These data contain the name of the IRB Member who made the initial determination, and in some cases the country of origin and gender of the claimant. I also use the commercial service Onomap to predict country of origin for each claimant, which I collapse to continent dummies.

For each judge, I collected information on the date and party of appointment. Appendix Table A1 contains summary statistics for the judges. 25% are female, and their dates of appointment range from 1982 to 2010. Since the Liberals held power for most of this time period, 72% of judges are Liberal appointees.[16] The average judge has 6.5 years of experience with a maximum of 28.

I exclude cases that were not perfected[17] or were unopposed by the government. I include only judges who decided cases in both the first and second round to improve comparability between the estimates of first- and second-round judge behavior.

3.5 Estimation

Fully nonparametric identification as outlined in Section 2.2 requires special regressors with large support conditional on judge assignment. This is a very high bar, and one that is not met by my empirical application. Instead, I parameterize the distributions of case strength r_i and judge errors $\widetilde{\varepsilon}_{ij1}$ and $\widetilde{\varepsilon}_{ij2}$, generating tractable analytic expressions for approval probabilities and allowing rapid estimation of the entire model by maximum likelihood. I begin by assuming that $\widetilde{\varepsilon}_{ijs}$ is mean-zero and normally distributed with standard deviation σ_{js} to be estimated as the measure of judge inconsistency. Larger σ_{js} corresponds to more inconsistency, ie a wider distribution of judge errors $\widetilde{\varepsilon}_{ijs}$. I additionally allow regressors W_{ijs} to affect errors, so

[15] Case files since 2006 are available at http://ccrweb.ca/en/2016-refugee-claim-data.

[16] The two main political parties in Canada are much closer ideologically than the major parties in the United States, as are the judges they appoint. There is less dissent within the legal community about the correct approach to statutory interpretation, although the Conservative party is generally more skeptical of refugee claims than the Liberal party.

[17] That is, those cases where all the paperwork was not filed on time and the appeal was automatically rejected.

$$\widetilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2), \;\; \sigma_{js}(W_{ijs}) = e^{\widetilde{\sigma}_{js} + W_{ijs}\psi_s} \tag{10}$$

As discussed in Section 3.2, the distribution of unobserved case strength for claimants at the Federal Court is likely right-skewed, since it is the distribution of individuals who were denied refugee status by government decision-makers. This captures the intuition that a relatively small number of high-case strength refugees are *not* initially granted status by the government. I therefore assume that the distribution is single-tailed; r_i is exponential-Pareto distributed (I show the distribution of r_i relative to the estimated parameters in Figure 3). I allow flexibility in the distribution of r_i across two dimensions. First, cases filed at different offices vary considerably in strength. This is partially because office is correlated with country of origin, but more closely related to varying levels of legal aid funding. Second, the government made changes to the decision process for initial refugee applications in 2002 (see Section 3.1 for more details). Combining these, I fix the distribution of r_i to have a scale and shape parameter of 1 for the largest office (Toronto) before 2002, and then separately estimate the scale and shape parameters for each office before and after 2002. In practice, I find almost no difference in the distribution of case quality between these time periods, suggesting that the institutional changes had little affect on which claimants were approved. However, the between-office variation is considerable.

In Appendix Section A2, I show that the probability of first-round approval with exponential-Pareto location parameter is x_m and scale parameter α is

$$\begin{aligned} P(r > \widetilde{\varepsilon}_{j1}(X_{ij1}, W_{ij1})) = \Phi\left[\frac{\ln(x_m) - \gamma_{j1} - X_{ij1}\beta_1}{\sigma_{j1}(W_{ij1})}\right] + \\ e^{\alpha(\ln(x_m) - \gamma_{j1} - X_{ij1}\beta_1) + \frac{\alpha^2 \sigma_{j1}(W_{ij1})^2}{2}} \left[1 - \Phi\left(\frac{\ln(x_m) - \gamma_{j1} - X_{ij1}\beta_1}{\sigma_{j1}(W_{ij1})} + \alpha\sigma_{j1}(W_{ij1})\right)\right] \end{aligned} \tag{11}$$

which can be calculated without resorting to computationally expensive numerical integration. Joint probabilities for first and second round approval are similar in spirit but more complicated. Since occasionally the same judge is assigned to the first- and second-round decision, I allow $\widetilde{\varepsilon}_{ij1}$ and $\widetilde{\varepsilon}_{ik2}$ to be correlated (with the correlation estimated as an additional parameter) whenever $j = k$. The full derivation and presentation is available in Appendix Section A2.

Full identification requires regressors X_{ijs} that shift judge thresholds γ_{js} but do not affect judge errors $\widetilde{\varepsilon}_{js}$. One ramification of this is that X_{ijs} must contain variables not in W_{ijs}, so that there is variation in X_{ijs} conditional on W_{ijs}.

One possible candidate for X_{ijs} is the order in which decisions are made. Chen, Moskowitz, and Shue (2016) show that US refugee judges are *less* likely to approve claimants when their previous case was granted asylum (this sort of negative correlation in decision-making also appears for baseball umpires and loan officers). In other contexts this might be a good choice; I unfortunately do not observe the decision order.

I instead use the timing of the decisions during the week, and of the second-round hearing during the day, as regressors. Danziger, Levav, and Avnaim-Pesso (2011) argue that when decision-makers make many decisions in a row, they become fatigued and are more likely to pick the default option. They study parole decisions in Israel and find that rejections become more likely just before lunch and revert to baseline levels immediately after the break. I follow them and use a dummy for the noon hearing (in the second round only) as a regressor.[18]

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

Identification using regressors leans on the assumption that variation in X_{ijs} does not affect errors. One might reasonably wonder whether this is true for the timing regressors. I address this issue in two ways. First, I estimate versions of the model where the excluded components of X_{ijs} are added in turn to W_{ijs} — I allow the regressors to directly affect the size of inconsistency. Since there are two regressors in the second round (dummies for the end of the week and a lunchtime hearing), I can then test whether $\sigma_{j2}(W_{ij2})$ varies with the regressors. In Section 4.5, I conduct these tests and find that the effect of the regressors on the error distribution is small and statistically insignificant.

Second, I estimate the model without relying on regressors. This version of the model leans more heavily on functional form assumptions to compare consistency for judges with different approval rates, but is a valuable robustness check. In Appendix Section A6, I present estimates for this model and show that all the results are qualitatively similar.

[18] There are two main reasons why one would expect denial to be the default option. First, most appeals are rejected. Second, because Canadian law requires judges show deference to governmental decision-makers, judges tend to see overruling IRB decisions as the exception rather than the rule.

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

4 Results

4.1 Randomization tests

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

The predictive power of judge assignment is low for the subsample of IRB-linked case files where I observe claimant characteristics. The coefficients from the regression of covariates on judge-level approval rates are all insignificant. Similarly to other examiner-effect contexts with random or quasi-random assignment, the F-statistics are small (about 1) but the test is sensitive enough to reject slightly more than half the time (Mueller-Smith, 2014).

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

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

4.2 Reduced form judge behavior

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

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

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

A more structural way to demonstrate the existence of a commonly-observed quality factor is to estimate the marginal treatment effect (MTE) of first-round approval on ultimate approval, instrumenting for first-round approval with judge assignment. I include this graph in Appendix Section A3, where I confirm that individuals marginally approved by more lenient judges in the first round are less likely to be approved in the second round.[19] The relationship is relatively weak and foreshadows the structural result that first-round judges have trouble selecting the claimants that will be successful in the second round; over the main mass of first-round approval rates the MTE declines from about 0.45 to 0.30.

Figure 1 demonstrates that there is substantial variation in judge propensity to approve refugee claims — in the language of the model, that there is variation in γ_{js}. For evidence on variation in σ_{js}, judges' ability to pick the highest-quality claimants, I turn to Figure 2. One source of identifying variation in the structural model comes from how often a first-stage judge's approved claimants are approved by a different judge in the next round. First-round judges with more ultimately successful claimants, the model implies, are better at picking high-r_i cases. Figure 2 displays reduced form evidence on the size of the variation in this ability. For each first-round judge, I take the mean

[19] The assumptions of MTE are violated when judges make errors, because individuals are no longer marginal with respect to any set of instruments. However, the graph can still be interpreted as an interesting descriptive exercise.

approval rate of her approved claimants in the second round. Panel A displays the histogram: a 10th percentile judge has 37% of his claimants ultimately approved; a 90th percentile judge 56%. Although it stands to reason that claimants approved by lenient judges in the first round would have a lower second-round success rate, the histogram changes little when I residualize out first-round approval rates and second-round judge approval rates (Panel B). In Panel C, I similarly calculate the second-round approval rates for claimants approved by each judge and plot them against the judges' first-round approval rates, residualizing out the second-round judge approval rate and shrinking the estimates toward the grand mean via Empirical Bayes to account for small cell sizes. The regression coefficient is negative and significant, but there is a large degree of cross-judge dispersion in second-round approval for each first-round approval average — the subsequent approval rate for judges approving about 15% of first-round applicants ranges from 38 to 48%. In other words, there is a lot of variation in the ability of judges to pick claimants who will be approved in the second round, even holding constant the share of claimants they approve.

4.3 Structural results

My baseline model includes as regressors X_{ijs} a dummy for a late-week decision and a dummy for whether the second-round hearing was heard at lunch. Thresholds γ_{js} vary by judge and round, and inconsistency σ_{js} varies by judge-round but not by any covariates (ie, W_{ijs} is empty). Case strength r_i is distributed as an exponential-Pareto, with different parameters in each office of case origination and before and after 2002, when there were changes to how the government made initial refugee determinations. I discuss these modeling decisions in detail in Section 3.5.

Figure 3 plots the distribution of judge-round thresholds γ_{js} and inconsistency 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 using the method of Delaigle, Hall, and Meister (2008). The coefficients are precisely estimated, so for most of the estimates this does not make a difference. For comparison I plot the distribution of case quality r_i in black.

In Panel A, the distribution of γ_{j1} is large relative to the distribution of refugee strength r_i, plotted in black. In Panel C, the distribution of second-round thresholds γ_{j2} is narrower and slightly smaller on average.[20] I interpret this as reflecting the larger amount of precedent available to judges in the second round relative to the first round. In the first round, no decisions are written up, making it difficult for judges to learn about the decisions other judges have made in similar situations. Conversely, nearly all second-round decisions are published as precedent. The correlation between round-specific γ_{js}'s is relatively high (0.46).

[20] The distribution of second-round γ_{j2} does not second-order stochastically dominate the distribution of γ_{j1}, but the standard deviation is smaller (0.39 vs. 0.74).

Panel B of Figure 3 shows the round-specific distributions of σ_{js}. The standard deviation of the first-round judge errors for the median judge is 1.08, which is large relative to the standard deviation of 1 for underlying case quality r_i. To contextualize the size of the inconsistency, I match pairs of first-round judges with approval rates within one percentage point. Using the two methods of quantifying disagreement I discuss in Section 2.3.3, I bound average disagreement at 3.6% of all cases. Using the more realistic assumption that cross-judge errors are uncorrelated, I calculate that the average pair of judges with the same approval rate disagrees on 13.2% of all cases. How big is this number? For *all* pairs of judges, the average disagreement rate is 23.1%, suggesting that inconsistency for similarly-severe judges is a larger contributor to how the justice system delivers different verdicts for the same claimant than cross-judge variation in severity.[21]

The distribution of σ_{j1} in Panel B is wide — some judges are much more consistent than others. A natural question is how changing the composition of judges would affect the overall disagreement rate. To answer this question, I simulate making the least-consistent half of judges as consistent as the median judge, adjusting thresholds γ_{j1} to keep each judges approval rate the same. I then re-estimate the disagreement rates, and find that bounded disagreement falls from 3.6 to 2.3%, and uncorrelated disagreement from 13.2 to 8% of all cases.[22] Policies that replaced the least consistent judges with average replacements would therefore have an important effect on the overall consistency of the justice system.

Estimates of the level of disagreement are affected by the overall approval rate — if two judges approve 1% each, they can disagree on at most 2% of the total cases. This is particularly relevant in this setting, where the first-round approval rate is only 14%. An alternative measurement of disagreement is to match pairs of judges with the same overall approval rate, then for each judge's approved cases calculate the share that would *not* be approved by the other judge. I conduct this exercise, and estimate that this measure of disagreement is bounded at 15.2%. Uncorrelated disagreement is a shocking 57.6%.

The presence of inconsistency has important implications for examiner-assignment research designs. This identification strategy uses random or quasi-random assignment of decision-makers to cases to generate random variation in a treatment, then studies the effect of treatment on outcomes. Prominent examples of the treatments studied include incarceration, patent receipt, and being placed in foster care (Bhuller et al., 2016; Gaulé, 2015; Doyle, 2008). The montonicity assumption in this context requires that all individuals are weakly more likely to be approved by a high-approval judge, and less likely to be approved by a low-approval judge (individuals for whom this does not hold are defiers in the Angrist et al. (1996) sense). The presence of inconsistency

[21] Even if all judges were perfectly consistent, cross-judge variation in approval rates alone implies that the average pair of judges disagrees on at least 8% of cases.

[22] Because I adjust the thresholds γ_{j1} to keep approval rates the same, these changes reflect only changes in the pair-specific disagreement rates, not changing pair composition.

implies violations of monotonicity. In Appendix Section A5 I examine the effect of inconsistency on IV and MTE estimates. Although this may be a setting where inconsistency is larger than other examiner-assignment contexts (suggesting large biases from inconsistency), the results are relatively reassuring. In a regression of second-round approval on first-round approval instrumented with judge assignment, even this relatively high level of inconsistency biases IV estimates by only 5%. More worryingly, inconsistency causes the estimated MTE to be considerably flatter than the true MTE.

In Panel D of Figure 3, the size of the distribution of second-round errors σ_2 is harder to interpret. Recall from Section 2.3.2 that the second-round error may contain an informational component $\mathcal{I}_{ij2}$. If this is the case, then a larger σ_{j2} may reflect better information-gathering abilities in the second-round hearing. Some suggestive evidence that information-gathering is an important part of second-round errors is in Appendix Figure A2, a scatter of judge-specific first- and second-round σ_{js}. The correlation is small and slightly negative (-0.043). It is reasonable to assume that the non-information components of first- and second-round errors are positively correlated, given that all the other measurable aspects of judge behavior (thresholds γ_j, overall approval rates) are correlated across rounds. This suggests there is a non-trivial information shock in the second round. The size of this shock is negatively correlated with the magnitude of first-round inconsistency — consistent first-round judges are also better at uncovering relevant information in the second-round hearing.[23,24] I take this as a further reason to focus on first-round consistency for the remainder of the paper.

Another interesting finding from Figure 3 is that most judge's estimated σ_{j2} is smaller than the σ_{j1} (though there is a long tail of large σ_{j1}). Why is this so striking? By construction, the information shock is orthogonal to other errors. This means that the distribution of judge errors $u_{ij} + e_{ij2}$ — which is necessarily smaller than $\widetilde{\varepsilon}_{ij2} = u_{ij} + e_{ij2} + \mathcal{I}_{ij2}$ after netting out informational shocks $\mathcal{I}_{ij2}$ — is smaller in the second round than in the first. This likely reflects that second-round judges take more time to make the decision, think more deeply, and usually write out a full decision explaining their reasoning.

First-round judges do a poor job of predicting which claimants will be successful in the second round; for most claimants the probability of second-round approval is higher than their probability of first-round approval.. Figure 4 shows how this works. For each r_i I calculate the first-round approval probability and the second-round approval probability conditional on first-round approval. I then plot them against each other. The figure shows that the median claimant has a 6% chance of first-round approval, but conditional on approval has a 12% chance of approval in the second round.

[23] Decomposing the first and second round residuals, we know that $\text{corr}(\text{var}(u_{ij}+e_{ij1}), \text{var}(u_{ij}+e_{ij2})+\text{var}(\mathcal{I}_{ij2})) \approx 0$, so if $\text{var}(u_{ij}+e_{ij1})$ and $\text{var}(u_{ij}+e_{ij2})$ are positively correlated that implies $\text{var}(u_{ij}+e_{ij1})$ and $\text{var}(\mathcal{I}_{ij2})$ are negatively correlated.

[24] Another (less believable but empirically indistinguishable) interpretation is that the informational-gathering component in the second round is small, but that the correlation in consistency across rounds is low.

This does not reflect selection, which is accounted for by conditioning on r. Instead, it shows that second-round decisions are unpredictable from the perspective of the first round, and even claimants with a relatively low quality factor r_i are sometimes approved in the second round. Furthermore, no one has a very high chance of approval: the 95^{th} percentile claimant has only a 62% chance of first-round approval and a 66% chance of second-round approval — a 41% chance of overall success at the Federal Court. The first-round approval rate is 14%; if the highest-r_i 14% was selected in the first round the overall approval rate would only climb to 8.5% (14% × 0.58) from its current value of 6%.

Integrating over the entire distribution of r, I find that on average 19.4% of claimants would be approved in a second-round hearing if they were approved in the first round. This is in contrast to the 6% of claimants who are approved under the current system. Although it may be surprising that the number of refugee appeals granted by the Federal Court would triple under this alternative decision-making process, the result is foreshadowed by the scatter plot of first-round approval rates by judge against the approval rates for that judges' approved claimants in the second round found in Figure 2. The most lenient judge approved 66.7% of claimants in the first round after partialling out the office and date of origin, and of those 27.4% were approved in the second round (partialling out the second-round judge, office and timing), bounding the overall approval rate in the absence of a first round at 18.3% (0.667 × 0.274). In terms of the policy implication, an important caveat is that judges might change their second-round behavior in unpredictable ways if all cases were automatically approved in the first round — one first-order effect would be that each judge would have 7 times more second-round cases to hear, which is likely not possible. This finding should therefore be interpreted more as a description of how well first-round judges can select the claimants that will be successful in the second round than a prediction of what would happen under a one-round system.

4.4 Relationship between structural parameters and reduced form statistics

In this section, I show how reduced-form moments translate into the structural parameters. In the model, the principal determinant of approval rates is judge thresholds γ. In Panel A of Figure 5, 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.

Recall from Section 2.2.1 that a main source of identification of the overall size of first-round judge inconsistency σ_{j1} is whether the approved claimants are subsequently approved in the second round. In Panel B, I adjust σ_{j1} away from its estimated values by multiplying each coefficient by a common factor. As inconsistency increases (σ_{j1} gets smaller), this dramatically reduces the second-round approval rate.

Panels C and D hew closer to the judge-randomization intuition of Section 2.2.1. Comparing two judges with the same first-round approval rate, the more consistent judge will have the higher approval rate in the second round for her approved claimants. In Panel C, I match judges with first-round approval rates within 1 percentage point of each other, then plot the difference in second-round approval rates against the difference in estimated σ_{j1}. As expected, judges with a higher second-round approval rate than their matched colleague have a lower estimated σ_{j1}, or in terms of the model are more consistent.

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

4.5 Decision timing as regressors

Identification requires that the case timing regressors affect judge thresholds γ_{js} but are not correlated with judge errors $\widetilde{\varepsilon}_{ijs}$ or case strength r_i. I explore whether the regressors are uncorrelated with case strength in Table 3. 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 judicial errors $\widetilde{\varepsilon}_{ijs}$ rather than case strength r_i or judge thresholds γ_{js}. If this were true, one possible implication is that the regressors would affect the distribution of errors. I test this directly in Appendix Table A3, where I include in turn the two decision timing regressors in W_{ijs}. In a nonparametric sense, the model is identified by an excluded regressor in each round that affects thresholds but not errors; since the noon-hearing regressor affects errors only in the second round these should be interpreted as tests on second-round identification. In line with the relevance tests in Table 3, both regressors have strong and statistically significant effects on γ_{js}. However, neither has a statistically significant affect on σ_{j2} — the log-log end-of-week coefficient is only 0.04 (SE=0.06) and the noon-hearing coefficient a very imprecisely estimated 0.37 (SE=1.29).

As an additional robustness check, in Appendix Section A6 I estimate the model without regressors and find that all the main results are qualitatively unchanged, albeit slightly less precise. Because this version of the model is identified only using judge randomization and functional form, the fact that the results are similar suggests that the effect of X_{ijs} on judicial errors is small.

4.6 Ideology versus observational errors

In Section 2.3.1, I discussed how judicial inconsistency can be decomposed into off-consensus ideological differences (inter-rater inconsistency) and pure observational errors (test-retest inconsistency). In this section, I provide some evidence on which factor is more important.

The Federal Court occasionally assigns the same judge to both the first- and second-round decision. I model this by allowing the observational errors $\widetilde{\varepsilon}_{ijs}$ to be correlated between rounds whenever the judge is the same in both rounds. The correlation is estimated to be positive and quite large, at 0.32 (SE=0.04). In a reduced-form sense, this reflects second-round justices disproportionately approving claimants that they approved in the first round, above and beyond what would be expected by their overall first- and second-round approval rates. One could interpret such a high correlation as resulting from either inter-rater inconsistency (a judge disproportionately values the strengths of the claimants case) or a common misreading of the facts of the case in both rounds, which I refer to as test-retest inconsistency. However, the size of the correlation implies that at least one of these factors is large. In Section 2.3.1 I describe a test of the relative size of these errors. If the correlation is caused by ideological inter-rater inconsistency, it is likely that pairs of judges share the same weighting of different aspects of cases. Then, in a regression of second-round approval on model-predicted likelihood of approval *and* judge-pair fixed effects, the fixed effects should add meaningful predictive power. One interpretation of this regression is as an overidentification test — the model is estimated under the assumption of no correlation between errors for different pairs of judges.

I implement the test in Table 4. 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. All specifications drop cases where the same judge made both the first- and second-round decisions.

Across columns, the coefficient on model estimates of approval probabilities is very close to one. However, in all specifications the F-stat for the joint test of judge-pair fixed effects is about

1, corresponding to p-values in the range of 0.30-0.60.[25] In other words, the judge-pair effects do not predict second-round approval beyond the model estimates. This suggests that there is in fact no correlation between errors $\widetilde{\varepsilon}_{ijs}$ for judge pairs, and that ideological errors are small relative to observational errors. As a descriptive analysis, I calculate the Empirical Bayes judge-pair means.[26] This confirms the F-stat result: the standard deviation of the judge-pair means is only 0.004 relative to a mean approval rate of 0.44.

The results in this section are a joint test of both the strength of the judge pair ideology correlations and the variance of the ideological errors; both must be large to generate ν_{jk}'s with detectable predictive power. It is unlikely, however, that the variance of the ideological errors u_{ik} is high but all the correlations are low, because that would require that u_{ik} is a very high-dimensional object. Refugee cases are fairly simple compare to other types of law: judges may differentially weight substantive and procedural aspects, as well as different types of refugee claims, but the complexity of the space is limited by the fact that the first-round decisions are made after reviewing the documentation from the original IRB decision, not holding full hearings. In other words, it is unlikely there are enough aspects of the cases that each judge could consistently weigh a different one of them more highly than all the other judges. I therefore take this as evidence that the judge errors $\widetilde{\varepsilon}_{ijs}$ are mostly composed of idiosyncratic observational errors, rather than differential weighting of different aspects of case strength.[27,28]

4.7 Judicial inconsistency, experience and workload

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

Table 5 presents models where I allow experience to enter the judge threshold γ_s (ie, in X_{ijs}) and the variance of the error, σ_{js} (ie, in W_{ijs}). I parameterize experience with an indicator for more than

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

[26] The Empirical Bayes means are the judge-pair residuals, shrunk towards the grand mean to account for measurement error.

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

[28] Another concern might be that the judge-pair test is low-powered because there are too many judge-pair cells. An alternative test that is similar in spirit but lower-dimensional is to use fixed effects defined by interactions of judge characteristics. Instead of testing whether knowing the exact identity of both judges has additional predictive, this test asks whether knowing the characteristics of the judges has additional predictive power. I implement this test for the gender, political party of appointment (Liberal or Conservative), and the native language of the judge (French or English, though most are bilingual), and find similar results to the judge-pair test — F-stats of about 1 and p-values in the 0.30-0.60 range.

one year of experience and with indicators for more than 1, 5, and 10 years of experience. Column 1 shows that first-round log inconsistency σ_{j1} shrinks by approximately 0.9 log points (60%) after the first year. In Column 2, I add additional indicators for more than 5 and more than 10 years of experience. The largest improvements are after the first year (0.77 log points), followed by further improvements after 5 (0.29 log points) and 10 years (0.54 log points). To put these numbers into context, I estimate that if all judges had less than one year of experience, pairs of judges with the same approval rate would disagree on 74% of approved cases. After one year of experience for all judges that declines to 54%; after a total of 5 years, 45%; and after 10 years 30%.[29]

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

Another factor that may affect judicial consistency is workload. Higher caseloads may reduce the amount of time judges can spend on cases, or make them work longer hours. To test this, I calculate monthly log workload as the number of leave cases a judge is assigned in a given month. Judges are also responsible for non-refugee cases, so this is an imperfect measure of workload. The model accounts for judge-specific consistency in σ_{js}, guaranteeing these estimates do *not* reflect time-invariant selection of more- or less-consistent judges into refugee work.[30] However, to the extent that judges with higher refugee caseloads may have lower non-refugee caseloads, these estimates are likely biased towards zero.

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

4.8 Judge inconsistency and expert opinion

The judge inconsistency parameters are related to readily-observable reduced form moments in the data, as well as with experience and workload in largely predictable ways. In this section, I explore whether they are also related to lawyer perceptions of judge ability. Higher degrees of correlation between model-based measures and expert opinion serve as a validation of the model, and suggest that it could be used as a diagnostic tool.

[29] In each counterfactual I adjust the thresholds γ_{j1} to keep overall approval rates the same. This means that the same judges are matched to each other in all counterfactuals.

[30] This is not a big fear. The Court claims they do not assign judges to cases or case types as a function of performance, fearing that this would result in challenges to the assignment procedure.

To measure expert opinion, I conducted an email survey of refugee lawyers who have appeared in refugee hearings at the Federal Court. I asked respondents to rate the judges with whom they had personal experience along dimensions analogous to the parameters of the model: how lenient is the judge to claimants (corresponding to judge threshold γ_{js}), and how consistent and predictable is the judge (corresponding to judge consistency σ_{js}). More details about the survey, including the question text and comparison of the respondents to the lawyer population, are in Appendix Section A4.

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

$$\widehat{C}_{j\ell} = \beta_0 + \beta_1 \text{Favorability}_{j\ell} + \beta_2 \text{Consistency}_{j\ell} + \eta_\ell + u_{j\ell} \tag{12}$$

where ℓ indexes lawyers and $\widehat{C}_j = \{\widehat{\gamma}_1, \widehat{\gamma}_2, \widehat{\sigma}_1\}$. I use model estimates that account for experience (which is highly predictive of behavior), and for each judge-respondent pair use the coefficient combinations reflecting experience at the time of their modal interaction. To account for estimation error in the model coefficients I use Hanushek's (1974) efficient estimator.[31] In Panel A, the dependent variable is the first-round $\widehat{\gamma}_1$. As expected, higher lawyer-reported favorability is associated with a lower threshold. The correlation is large but imprecise in the first round; in the right-most preferred specification one SD higher favorability corresponds to a 0.19 lower γ_1, which is about 0.21 SD of the cross-judge distribution of γ_1. Panel B displays the relationship between γ_2 and the survey measures. The relationship is stronger than with γ_1; adding one SD of predicted favorability decreases γ_2 by 0.42, or 0.42 SDs of the judge distribution. This may be because second-round judge behavior is more salient than first-round behavior for lawyers, since they appear in front of the judge only in the second round.

Finally, Panel C shows the relationship between reported judge characteristics and σ_1. Reported consistency is negatively related with the model estimate of σ_1; across specifications one extra SD of consistency translates to between 0.16 to 0.22 lower σ_1, or about 0.1 SD of the cross-judge distribution. In other words, the model and the judge survey select the same judges as being more consistent, suggesting that the structural model is picking up true variation in judge ability to assess case strength and use common standards.[32]

[31] Hanushek's two-step method exploits knowledge of the standard error of the dependent variable C_j (ie, the model coefficients) to construct observation-level estimates of the variance of the residual u_j. The second step reweights observations by the inverse standard deviation of the residual.

[32] Because second-round σ_2 should not be interpreted as reflecting judge consistency, I do not include it in the table. However, consistent with it being partially correlated with true judge consistency, I find that σ_{j2} is negatively but insignificantly correlated with surveyed inconsistency.

4.9 Judge selection reform and judge consistency

In 1988, the government enacted an important reform to how it selects judges. The goal of the reform was to make it harder for the party in power to appoint unqualified party supporters. As I detail in Section 3.3, the limited evidence available suggests that the policy change reduced the number of new judges with ties to the ruling party. In this section, I provide evidence that the reform was also successful in reducing judge inconsistency σ_{j1}.

Table 7 presents a regression of $\widehat{\sigma}_{j1}$ on a dummy for whether the judge was appointed before the reform. I weight the regressions to account for estimation error in the dependent variable (Hanushek, 1974), and in my preferred specifications control for judge gender and party of appointment. Because the reform took place seven years before the start of my sample, the pre-reform judges are mechanically more experienced. More experienced judges are more consistent (have lower σ_{j1}), so this likely works against finding that the reform improved judge consistency.[33]

I show results for a baseline model that does not control for experience, and controlling for experience with categorical variables for more than 1, 5 and 10 years of experience (for approximate comparability, I adjust all coefficients to the median experience of 6 years). The first three columns show that average consistency improved by 0.34 after the reform, a 26% reduction (the estimate is just shy of statistical significance, with a p-value of 0.12. This does not appear to be related to the change in party that occurred just after the reform — Column 3 shows that party has no affect on large of statistically significant effect on consistency. In Columns 4-6, the the effect is larger once the model properly accounts for differing experience among pre- and post-reform judges, with σ_{j1} dropping by 1.7 points (relative to a pre-reform mean of 2.2) for judges appointed after the reform.[34]

Both of these affects are large. The estimates from my preferred right-most model imply a pre-reform uncorrelated disagreement rate of 18.3% for pairs of judges with the same approval rate (recall the baseline estimate is 13.2%). If all judges were appointed post-reform, that drops to 6.6%.[35] The strength of the effect speaks to the size of the reform, which materially restricted the minister's options. Government data shows that the Judicial Advisory Councils approve only 40% of applicants; ostensibly some of the rejected candidates would otherwise have been appointed.

[33] Alternatively, if high-consistency judges are more likely to be promoted to a higher court, then pre-reform consistent judges might not be observed in my data. This would mechanically make the pre-reform judges look less consistent. Although about 20% of the justices are promoted in seven or fewer years, there is not a strong or statistically significant correlation between promotion and estimated consistency — judges who are eventually promoted have a log σ_{j1} 0.15 lower (standard error 0.207) than non-promoted judges.

[34] In Table A7, I estimate an identical table using second-round inconsistency σ_2 as the dependent variable. Because σ_2 may partially reflect skill at gathering information in hearings, the predicted effect is ambiguous. In all but one specification, I find statistically insignificant results. This suggests either that post-reform judges made fewer errors only in the first round (which is implausible), or that information acquisition is an important part of second-round errors.

[35] To calculate these numbers I scale each σ_{j1} by a common factor so that the mean approval matches the before and after mean. This maintains a comparable amount of cross-judge variation in σ_{j1}, which is an important contributor to inconsistency.

Interestingly the effect is not driven by changes in judge leniency. In Appendix Table A6, I show that judges appointed after the reform approved a similar share of first-round claimants (an insignificant 5 percentage points more). More directly, a version of Table 7 that controls for changes in judge standards γ shows similar results, with an average σ_{j1} 0.98 smaller for judges appointed after the reform (SE=0.32).

The table also shows that there is no significant or substantial difference in judicial error between the two parties. Given the relative similarity in judicial philosophies between liberal and conservative judges in Canada, this finding is not particularly surprising. The Federal Court is a prestigious appointment, and so governments are unlikely to be constrained by supply limitations. Male judges are slightly (and marginally statistically significantly) more consistent than female judges, though this difference is dwarfed by both the gains to experience and the post-reform effect.

4.10 Optimal judge allocation

The Court assigns cases to judges taking into account only their availability, not their behavior in previous cases. In this section, I show how the Court could optimize judge allocation to minimize caseload while maintaining the same standards.[36]

Second-round decisions are much more costly to the court than first-round decisions. Instead of reading documents from the IRB's initial determination, a second-round decision entails a full hearing in front of the opposing lawyers, time to prepare for the hearing and time to write the decision — about ten times as long as a first-round decision. The Court could minimize workload while approving the same number of total claimants by reducing the number of first-round acceptances and approving all second-round claimants. To some extent, they are already pursuing this strategy — as I discuss in Appendix Section A3, the marginal claimants for most first-round judges have a relatively high chance of second-round approval (30-45%).

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

Under this problem, there are three ways to minimize caseload. First, judges should be reallocated to rounds where they make more consistent decisions. Second, first-round judges should be made more strict to improve the posterior case quality of claimants approved in the first round

[36] Alternatively, I could maximize acceptance rates while maintaining the same standards and keeping workload no higher than in baseline. This problem is almost symmetric, and for given model estimates the potential cost savings holding acceptance fixed are always close to the percentage increase in refugees holding costs fixed.

and decrease the number of second-round decisions. Third, second-round judges should approve a higher share of cases so that the overall approval remains the same given lower first-round approval rates.

In Figure 6, I conduct exactly this maximization. I find that overall workload would be reduced by 17.5% (or 28,000 hours), amounting to savings of approximately \$4.4 million in judge salaries alone over the study period. This counterfactual poicy would also save staff time and allow claimants to receive their ultimate decision faster. The figure demonstrates the second two kinds of savings, but not the first. To summarize how the re-assignment procedure works, I present histograms of the baseline judge coefficients by round as well as histograms that have been reweighted to reflect the distribution of coefficients after optimization. The average first-round threshold γ_{j1} for optimally-assigned judges is higher (Panel A), meaning that fewer cases will be approved in the first round but case strength conditional on first-round approval will be higher. Conversely, judges in the second round are much more lenient, as evinced by the lower thresholds γ_{j2} (Panel C). However, in Panel B and D the change in the overall distribution of consistency is much less dramatic — only the most inconsistent judge-rounds are eliminated. I interpret this to mean that judge thresholds γ_{js} are a stronger driver of who is selected, but that knowledge of judge-specific consistency has an important role to play in allowing the researcher to discipline the selection process.

The problem as I've described it takes the posterior distribution of r_i as the relevant measure of quality. Implicitly, this assumes that the second-round error is all judge error. As I discuss in Section 2.3.1, it may also reflect additional information gained in the second-round hearing. In that case, the relevant measure of quality is $r_i + \mathcal{I}_{ij2}$. I do not directly estimate the distribution of of the information shock $\mathcal{I}_{ij2}$, so cannot perfectly condition on the posterior (I estimate the distribution of inconsistency plus information shock, $u_{ij} + e_{ij2} + \mathcal{I}_{ij2}$). However, under the assumption that the second-round error is *all* information, I can minimize workload so that the posterior of $r_i + \mathcal{I}_{i2}$ first-order stochastically dominates the baseline distribution. Under this specification, the allocation of judges is highly correlated with the baseline optimization (0.58), and workload is reduced by 16% rather than 17.5%. It also satisfies the constraints of the baseline allocation, so a cautious planner could implement the second design and enjoy most of the gains of judge reallocation.

5 Conclusion

Much research has focused on non-relevant factors that affect judge behavior: the decision in the previous case (Chen, Moskowitz, and Shue, 2016), the outcome of a college football game (Eren and Mocan, 2016), or the timing of the hearing relative to lunch (Danziger et al., 2011). The existence of these phenomena suggests that the same defendant could be convicted by one judge and acquitted by another, even when both judges have the same overall incarceration rate. In this paper I develop techniques to quantify the prevalence of this type of inconsistency.

I begin with a simple model where judges approve all candidates with a case strength larger than a judge-specific threshold. Judges observe case strength with some error, which generates inconsistencies across judges in which claimants they approve, even for judges who approve the same share of cases. I show that this model is identified in two-stage judicial processes by a combination of cross-judge comparisons (for example, more consistent first-stage judges are more likely to have their approved claimants approved in turn by the second round judge) and regressors that shift judge thresholds without affecting errors. Under parametric assumptions it can be tractably estimated.

I implement the model using data on judicial review of initially-denied refugee claims at the Federal Court of Canada. Although the justices of the Federal Court are experts in refugee cases, I uncover relatively high levels of inconsistency. For first-round judges who approve the same share of cases, I estimate they disagree on 13.2% of cases, and bound disagreement to at least 3.6% of cases. Disagreement is even higher among the cases they approve, where judges disagree on 57.6% on average (and are bounded at 15.2%). Overidentification tests suggest that most disagreement arises from idiosyncratic observational errors, rather than permanent differences between judges in which aspects of cases they think are most relevant.

Cross-judge variation in inconsistency is large. I validate the measured variation against a survey that solicited estimates of judicial characteristics from refugee lawyers who had appeared in front of the judges. Judicial consistency improves dramatically after the first year, and continues to improve (albeit at a slower rate) for at least the first ten years of experience. Inexperienced judges — but not experienced ones — are more consistent when they have a smaller workload. A reform in the late 1980s designed to stop the government from appointing unqualified party supporters dramatically improved judicial consistency, suggesting that well-designed judge selection processes can indeed improve court outcomes. Because my model generates measures of the posterior distribution case quality of approved claimants, I construct a counterfactual allocation of judges to cases that first-order improves on the posterior distribution while reducing judge workload. I estimate that the optimal policy would reduce judge hours by 18%, saving at least $4.4 million over the study period.

It is unclear how general this result is. By construction the Court's caseload is difficult, consisting of initially-denied refugee claimants who appeal the decision, and the lack of precedent likely increases inconsistency. In the Appendix, I show that this level of inconsistency implies relatively large levels of bias in MTE estimates using judge-assignment instruments, but not for linear IV estimates. Future work should determine whether the results hold in criminal courts and other decision-making institutions such as the Social Security Administration, and once the econometrician can condition on type of case and other covariates. If the level of inconsistency I uncover is also present in other contexts, it would introduce bias into estimates of the effect of incarceration (Aizer and Doyle, 2013; Mueller-Smith, 2014), SSDI receipt (Maestas et al., 2012), and patent receipt

(Gaulé, 2015) recovered from examiner-assignment IV designs.

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

References

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.

Aizer, A. and J. J. Doyle (2013): "Juvenile incarceration, human capital and future crime: Evidence from randomly-assigned judges," Tech. rep., National Bureau of Economic Research.

Alesina, A. F. and E. L. Ferrara (2011): "A test of racial bias in capital sentencing," Tech. rep., National Bureau of Economic Research.

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.

Anwar, S. and H. Fang (2006): "An alternative test of racial prejudice in motor vehicle searches: Theory and evidence," *The American economic review*, 96, 127–151.

Bhuller, M., G. B. Dahl, K. V. Løken, and M. Mogstad (2016): "Incarceration, recidivism and employment," Tech. rep., National Bureau of Economic Research.

Bureau of Justice Statistics (2006): "Examining the Work of State Courts," .

Canes-Wrone, B., T. S. Clark, and J. P. Kelly (2014): "Judicial selection and death penalty decisions," *American Political Science Review*, 108, 23–39.

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.

Chandra, A. and D. O. Staiger (2011): "Expertise, underuse, and overuse in healthcare," .

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, X., J. Heckman, and E. Vytlacil (1999): "Identification and N Efficient Estimation of Semiparametric Panel Data Models with Binary Dependent Variables and a Latent Factor," *Cahier de recherche.*

Chen, X., J. J. Heckman, E. Vytlacil, et al. (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.

Coase, R. H. (1960): "The problem of social cost," *The Journal of Law and Economics*, 3, 1–44.

Dahl, G. B., A. R. Kostol, and M. Mogstad (2013): "Family welfare cultures," Tech. rep., National Bureau of Economic Research.

Danziger, S., J. Levav, and L. Avnaim-Pesso (2011): "Extraneous factors in judicial decisions," *Proceedings of the National Academy of Sciences*, 108, 6889–6892.

Dauvergne, C. (2003): "Evaluating Canada's new Immigration and Refugee Protection Act in its global context," *Alta. L. Rev.*, 41, 725.

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.

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.

Epstein, L., W. M. Landes, and R. A. Posner (2013): *The Behavior of Federal Judges: A Theoretical and Empirical Study of Rational Choice*, Harvard University Press.

Eren, O. and N. Mocan (2016): "Emotional judges and unlucky juveniles," Tech. rep., National Bureau of Economic Research.

Fischman, J. B. (2008): "Decision-making under a norm of consensus: A structural analysis of three-judge panels," in *1st Annual Conference on Empirical Legal Studies Paper*.

——— (2013): "Measuring Inconsistency, Indeterminacy, and Error in Adjudication," *American Law and Economics Review*, 16, 40–85.

Frakes, M. D. and M. F. Wasserman (2014): "Is the time allocated to review patent applications inducing examiners to grant invalid patents?: Evidence from micro-level application data," *Review of Economics and Statistics*.

Gaulé, P. (2015): "Patents and the success of venture-capital backed startups: Using examiner assignment to estimate causal effects," .

Glaze, L. E. and E. Parks (2011): "Correctional populations in the United States, 2011," *Population*, 6, 8.

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. and E. Vytlacil (2005): "Structural equations, treatment effects, and econometric policy evaluation," *Econometrica*, 73, 669–738.

Keung, N. (2011): "Refugee board member with zero acceptance rate chastised," *The Toronto Star*.

Klein, T. J. (2010): "Heterogeneous treatment effects: Instrumental variables without monotonicity?" *Journal of Econometrics*, 155, 99–116.

Loeffler, C. E. (2013): "Does imprisonment alter the life course? Evidence on crime and employment from a natural experiment," *Criminology*, 51, 137–166.

Maestas, N., K. J. Mullen, and A. Strand (2012): "Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt," .

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. (2014): "The criminal and labor market impacts of incarceration," *Unpublished Working Paper*.

Norris, S. and M. Pecenco (2017): "The Intergenerational Effects of Incarceration: Evidence from 20th Century Iowa," .

Partridge, A. and W. B. Eldridge (1974): *The Second Circuit sentencing study: A report to the judges of the Second Circuit*, Federal Judicial Center.

Pew (2012): "Assessing the Representativeness of Public Opinion Surveys," Tech. rep., Pew Research Center.

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.

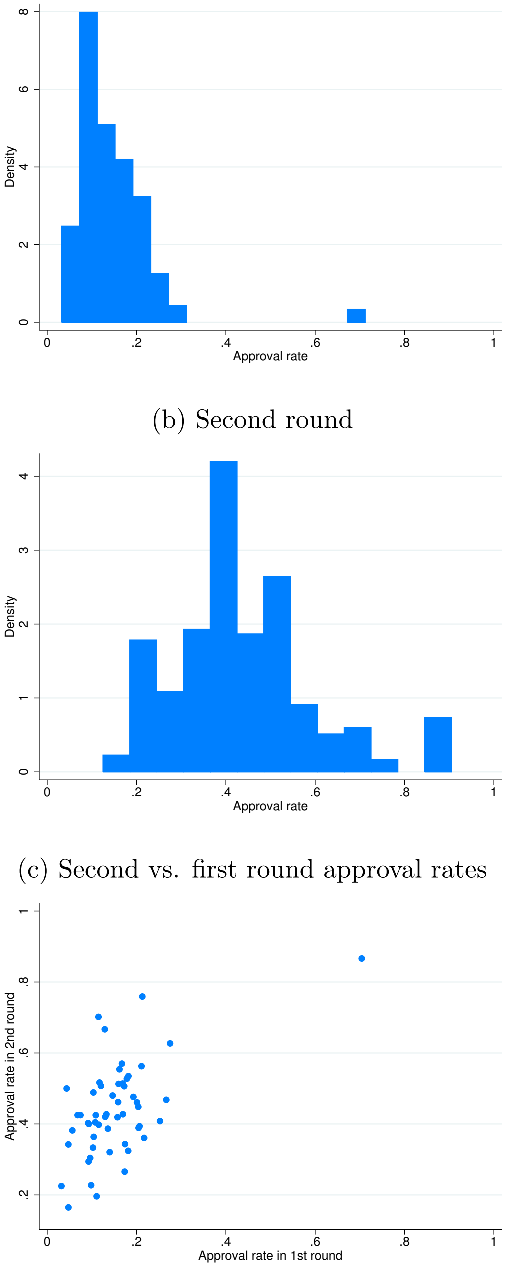
Russell, P. H. and J. S. Ziegel (1991): "Federal Judicial Appointments: An Appraisal of the First Mulroney Government's Appointments and the New Judicial Advisory Committees," *The University of Toronto Law Journal*, 41, 4–37.

Sah, R. K. and J. E. Stiglitz (1986): "The architecture of economic systems: Hierarchies and polyarchies," *The American Economic Review*, 716–727.

Shayo, M. and A. Zussman (2010): "Judicial ingroup bias in the shadow of terrorism," *Quarterly Journal of Economics, Forthcoming.*

6 Figures

Figure 1: Approval rates by judge



(a) First round

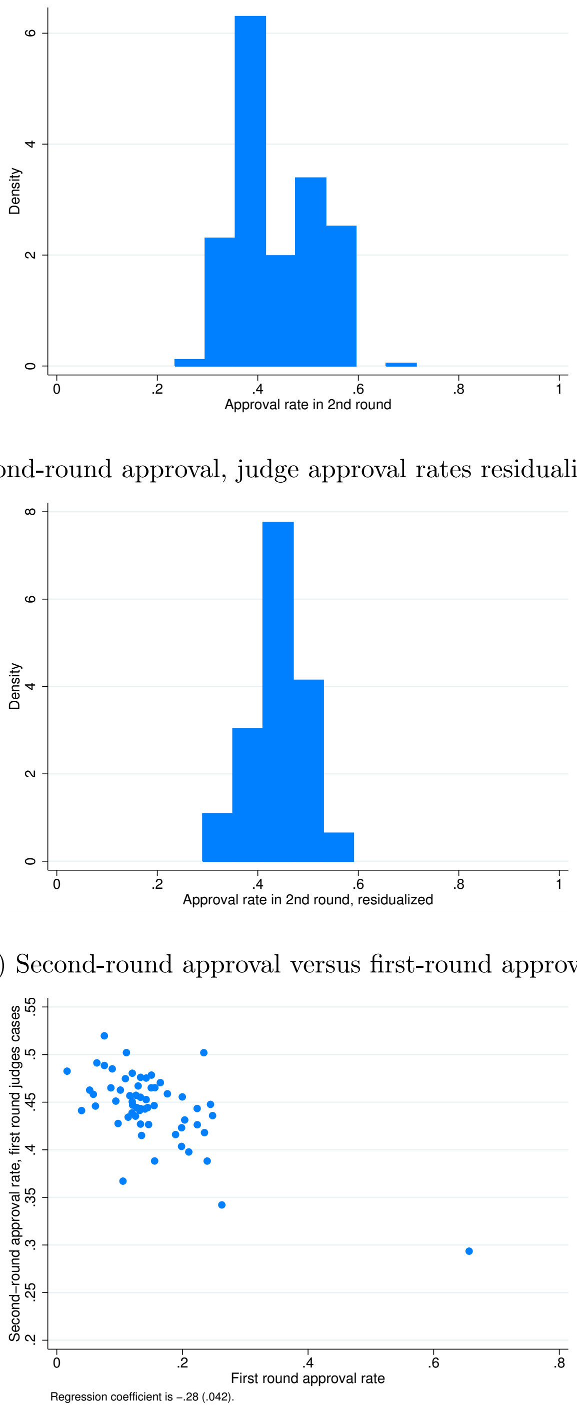
(b) Second round

(c) Second vs. first round approval rates

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

Figure 2: Second-round approval by first-round judge

(a) Second-round approval



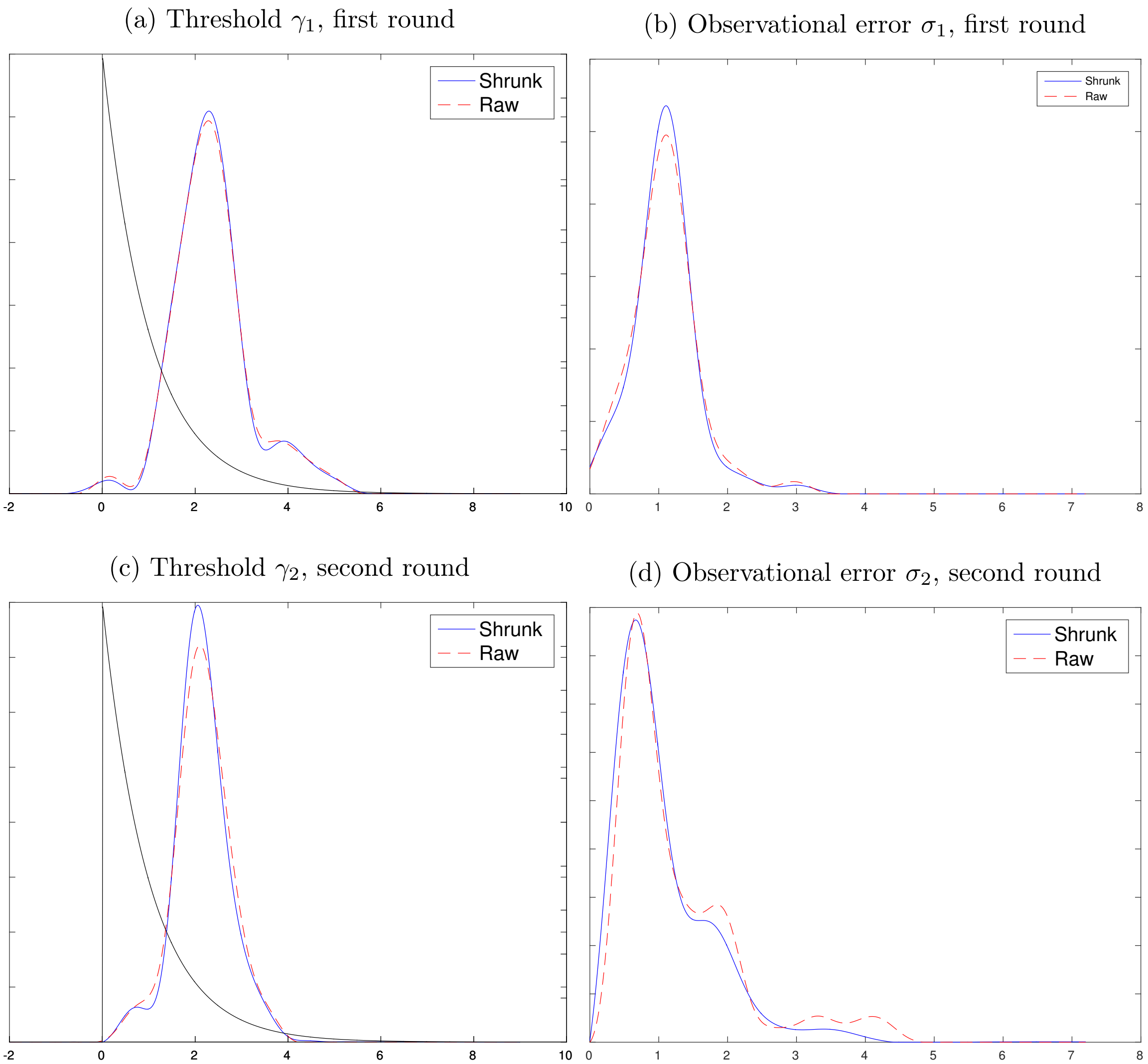
(b) Second-round approval, judge approval rates residualized out

(c) Second-round approval versus first-round approval

Regression coefficient is −.28 (.042).

Panel A contains a histogram of second-round approval rates for the cases approved by the first-round judge. Higher approval rates suggest that the first-round judge did a better job of selecting claimants with a high probability of success in the second round. Panel B residualizes out first- and second-round judge approval rates. Panel C shows second round approval rates for the claimants approved by each first round judge plotted against the judge's first-round approval rates, with second-round judge approval rates residualized out and means shrunk towards the grand mean via Empirical Bayes to account for measurement error.

Figure 3: Distribution of judge coefficients


(a) Threshold γ_1, first round

(b) Observational error σ_1, first round

(c) Threshold γ_2, second round

(d) Observational error σ_2, second round

This figure presents coefficient estimates for the decision model $\mathbb{1}[r_i > \gamma_{js} + X_{ijs}\beta_s + \widetilde{\varepsilon}_{ijs}]$, $\widetilde{\varepsilon}_{ijs} \sim \mathcal{N}(0, \sigma^2_{js})$. All models include controls for time/date of decision in β_s, and allow the parameters of the Pareto distribution of r_i to vary flexibly by office of origination as well as after 2002 relative to before 2002, when there were legislative changes that may have impacted the distribution of case quality. Each panel contains the density of the raw and shrunken estimates of the judge-round specific thresholds γ_1 and γ_2, and inconsistency σ_1 and σ_2. Black line is density of case quality r. Shrunken estimates recovered via deconvolution of estimates accounting for coefficient-specific standard errors, clustered at the level of the first round judge (Delaigle and Meister, 2008).

Figure 4: Model estimates of first- versus second-round approval

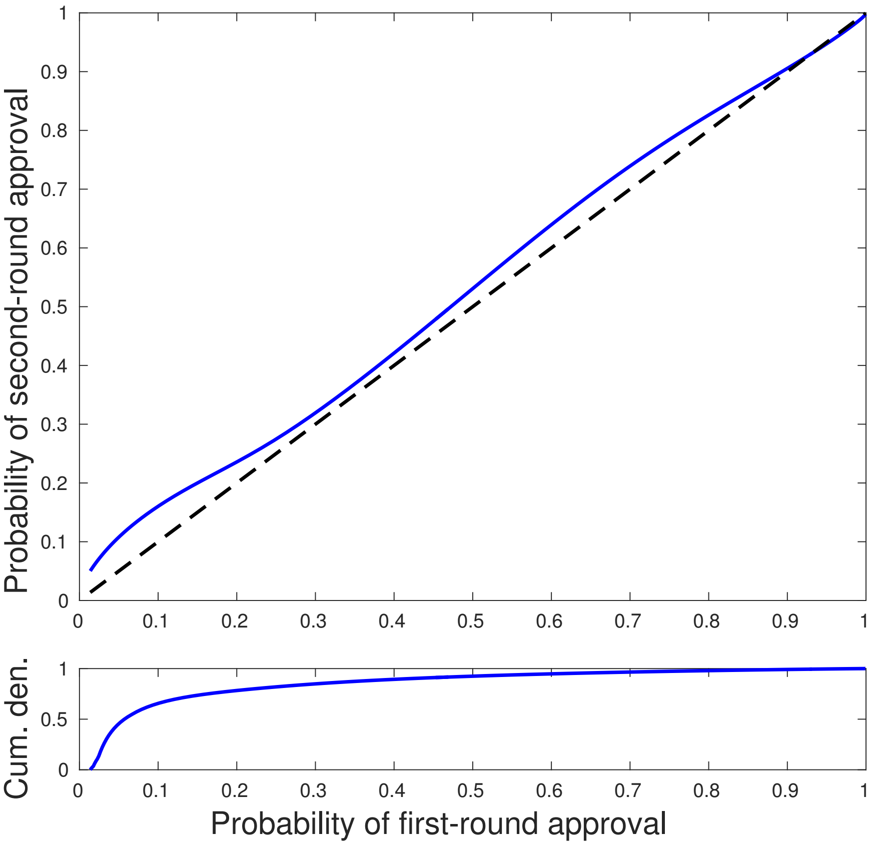
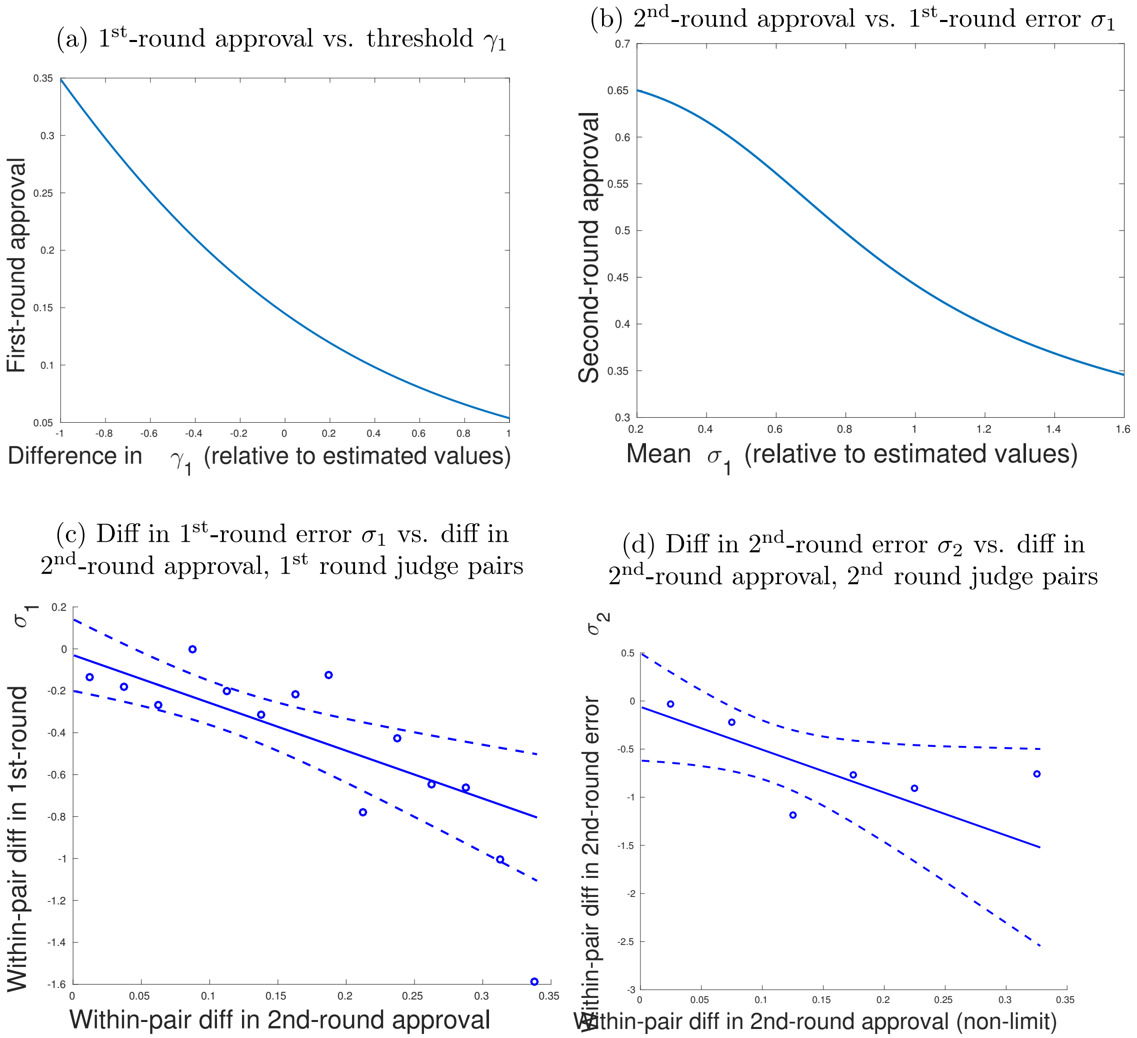


Figure plots first-round approval probability against second-round approval probability conditional on first-round approval for each value of case strength r_i. Secondary graph displays cumulative density of first-round approval. Black dotted comparison line marks out 45°.

Figure 5: Identification intuition



(a) 1^{st}-round approval vs. threshold γ_1

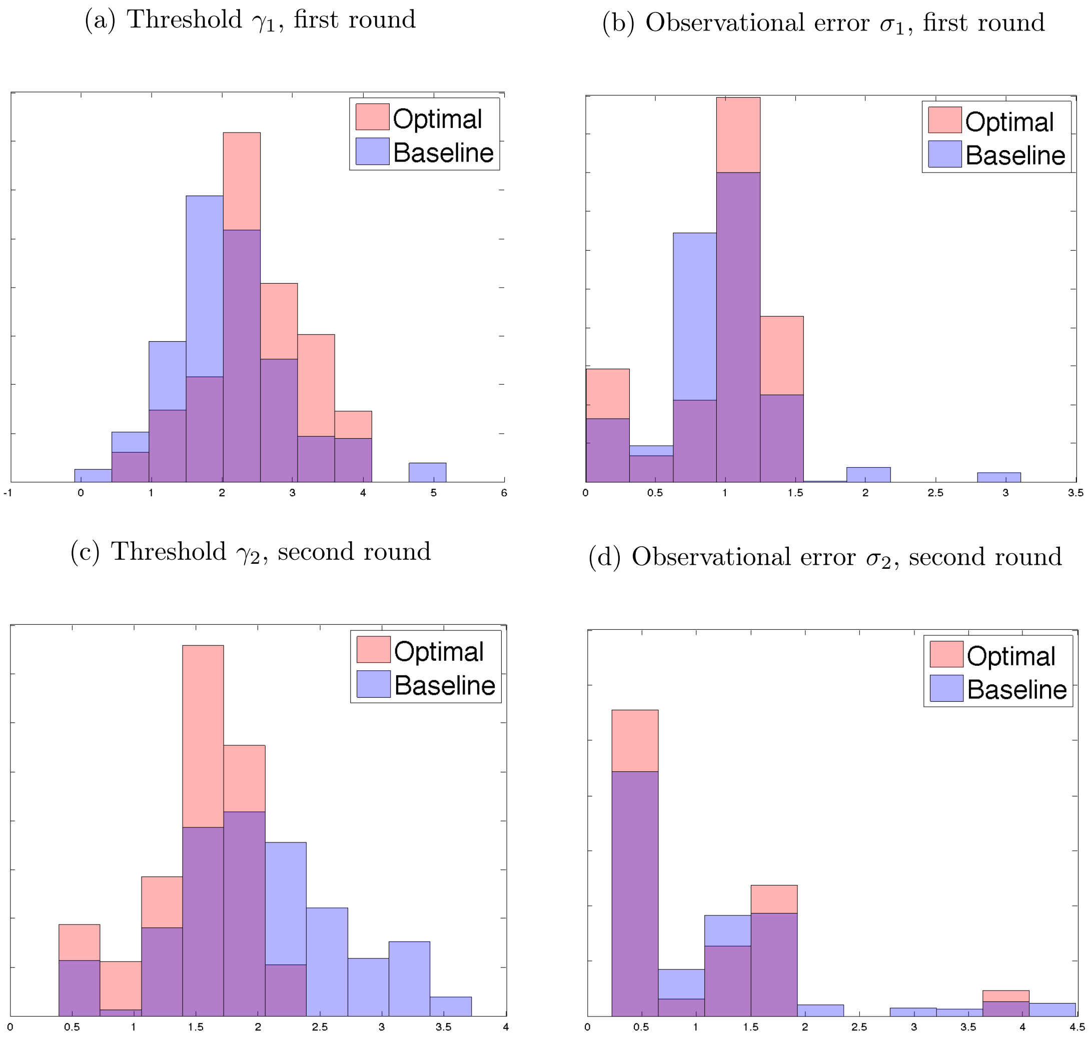
(b) 2^{nd}-round approval vs. 1^{st}-round error σ_1

(c) Diff in 1^{st}-round error σ_1 vs. diff in 2^{nd}-round approval, 1^{st} round judge pairs

(d) Diff in 2^{nd}-round error σ_2 vs. diff in 2^{nd}-round approval, 2^{nd} round judge pairs

Panel A contains model estimates of the first-round approval probability as a function of deviation of threshold γ_1. Panel B contains model estimates of second-round approval as a function of mean first-round error relative to the estimated value — higher errors make second-round approval less likely. In Panel C, I match pairs of judges with similar first-round approval rates (within 1 percentage point), then display difference in model-estimated judge errors $\widehat{\sigma}_1$. In Panel D, 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 Section 2.2.1 for more details.

Figure 6: Optimal allocation of judges


(a) Threshold γ_1, first round

(b) Observational error σ_1, first round

(c) Threshold γ_2, second round

(d) Observational error σ_2, second round

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

7 Tables

Table 1: Randomization

	Male	Africa	Asia	South America	IRB mean approval	Predicted approval	1st-round mean approval
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: First round judges							
First-round mean approval rate	-0.059	-0.019	-0.004	0.050	0.016	-0.000	
	(0.036)	(0.052)	(0.039)	(0.053)	(0.025)	(0.003)	
F-stat	0.90	2.17	1.44	2.25	3.75	2.08	
Prob	0.67	0.00	0.02	0.00	0.00	0.00	
Observations	19,436	19,436	19,436	19,436	19,436	19,436	
Panel B: Second round judges							
Second-round mean approval rate	-0.011	0.059	0.015	0.005	0.031	0.011	0.008
	(0.062)	(0.036)	(0.055)	(0.079)	(0.021)	(0.009)	(0.021)
F-stat	1.06	0.80	0.98	1.63	1.63	1.71	2.46
Prob	0.36	0.82	0.51	0.00	0.01	0.00	0.00
Observations	3,414	3,414	3,414	3,414	3,414	3,414	3,414

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. All regressions include office X pre-2002 fixed effects to account for cross-office differences in case strength and changes in government policy in 2002. Standard errors clustered at the judge level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Second-round approval on mean approval ratee of first-round judge

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

Standard errors clustered by second-round judge. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Placebo tests and relevance for regressors, with judge fixed effects

	Predicted approval		Actual approval	
	(1)	(2)	(3)	(4)
Panel A: First round				
End of week	0.000	0.000	-0.008***	-0.007***
	(0.000)	(0.000)	(0.002)	(0.002)
Observations	58604	58604	58604	58604
Panel B: Second round				
End of week	0.001	0.001	-0.022*	-0.022*
	(0.001)	(0.001)	(0.012)	(0.012)
Noon hearing	-0.001	-0.001	-0.078***	-0.075***
	(0.001)	(0.001)	(0.022)	(0.023)
Controls	No	Yes	No	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 4: Second-round outcome on model approval probability and judge-pair FEs

	Judge-pair round FEs		Judge-pair FEs	
	(1)	(2)	(3)	(4)
Model approval probability	0.945***	0.938***	0.967***	0.962***
	(0.146)	(0.169)	(0.0463)	(0.0470)
Model controls	No	Yes	No	Yes
Mean approval	0.44	0.44	0.44	0.44
F-stat for judge pairs	1.01	1.02	0.99	0.99
P-value	0.350	0.344	0.619	0.615
Bootstrap p-value	0.629	0.621	0.490	0.495
SD of judge-pair EB means	0.004	0.004	0.004	0.004
Observations	8,196	8,196	8,196	8,196

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 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 5: First-round judge consistency by experience and workload

	(1)	(2)	(3)	(4)	(5)
			Coefficients ψ affecting judge inconsistency σ_1		
Experience > 1 year	-0.899^{***}	-0.774^{***}		-0.337^{***}	-0.460^{***}
	(0.091)	(0.096)		(0.119)	(0.144)
Experience > 5 years		-0.290^{***}		0.047	-0.053
		(0.046)		(0.080)	(0.383)
Experience > 10 years		-0.536^{***}		-0.476^{***}	-0.509^{***}
		(0.175)		(0.100)	(0.113)
Log caseload			0.197^{***}	0.239^{***}	
			(0.009)	(0.023)	
Log caseload ($\leq$ 5 yrs exp)					0.179^{***}
					(0.040)
Log caseload (> 5 yrs exp)					0.062
					(0.100)
Second-round experience control	Yes	Yes	No	Yes	Yes

Reports coefficients for decision model $\mathbb{1}[r_i > \gamma_{js} + X_{ijs}\beta_s + \widetilde{\varepsilon}_{ijs}(W_{ijs})]$, $\widetilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\widetilde{\sigma}_{js} + W_{ijs}\psi_s)$. 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. Standard errors clustered at the level of the first stage judge. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Model coefficients on survey responses

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Threshold γ_1 (mean=2.26, SD=.87)						
Favorability, SD	-0.144	-0.237***			-0.055	-0.186*
	(0.104)	(0.086)			(0.114)	(0.102)
Consistency, SD			-0.316***	-0.212***	-0.304***	-0.134*
			(0.046)	(0.055)	(0.052)	(0.071)
Respondent FE	No	Yes	No	Yes	No	Yes
Observations	182	182	182	182	182	182
Panel B: Threshold γ_2 (mean=2.08, SD=.99)						
Favorability, SD	-0.285***	-0.409***			-0.265***	-0.416***
	(0.094)	(0.080)			(0.086)	(0.090)
Consistency, SD			-0.134*	-0.137	-0.070	0.019
			(0.069)	(0.088)	(0.058)	(0.078)
Respondent FE	No	Yes	No	Yes	No	Yes
Observations	182	182	182	182	182	182
Panel C: Inconsistency σ_1 (mean=1.89, SD=2.17)						
Favorability, SD	0.082	0.020			0.152***	0.092
	(0.060)	(0.099)			(0.044)	(0.094)
Consistency, SD			-0.184***	-0.163***	-0.224***	-0.194***
			(0.047)	(0.056)	(0.049)	(0.061)
Respondent FE	No	Yes	No	Yes	No	Yes
Observations	182	182	182	182	182	182

Reports linear regressions of model coefficients on survey responses, estimated with Hanushek (1974) correction for estimated dependent variable. Decision model is $\mathbb{1}[r_i > \gamma_{js} + X_{ijs}\beta_s + \widetilde{\varepsilon}_{ijs}(W_{ijs})]$, $\widetilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\widetilde{\sigma}_{js} + W_{ijs}\psi_s)$. 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. I adjust for judge experience in β_s and ψ_s. Model standard errors clustered at the level of the first stage judge, linear standard errors at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: Inconsistency before and after judge selection reform

	Baseline			Experience control in σ_1		
	(1)	(2)	(3)	(4)	(5)	(6)
After reform (=1)	-0.215	-0.332	-0.340	-1.740***	-1.785***	-1.752***
	(0.151)	(0.207)	(0.214)	(0.594)	(0.629)	(0.647)
Liberal appointee (=1)			-0.0133			-0.108
			(0.106)			(0.109)
Male judge (=1)			-0.0873			-0.210*
			(0.101)			(0.122)
Year appointed	No	Yes	Yes	No	Yes	Yes
Pre-reform mean	1.29	1.29	1.29	2.20	2.20	2.20
N judges	53	53	53	53	53	53

Estimated with Hanushek (1974) correction for estimated dependent variable. Dependent variable is consistency σ_{j1}, which is estimated from decision model $\mathbb{1}[r_i > X_{ijs}\beta_s + \gamma_{js} + \widetilde{\varepsilon}_{ijs}(W_{ijs})]$, $\widetilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\widetilde{\sigma}_{js} + W_{ijs}\psi_s)$. 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix for *Judicial Errors: Evidence from Refugee Appeals*

A1 Details on the use of regressors for identification

With a small change of notation, the main model of Section 2 can be recast as a single-spell duration model (Chen et al., 1999), where the "duration" is the amount of time until a judge rejects the applicant's case (duration is capped at 2). Equation 1 from the main text sets out the problem as identifying the parameters of the choice model where approval in each stage s occurs if

$$r_i > \varepsilon_{ijs}(X_{ijs}, W_{ijs}) = \gamma_{js} + X_{ijs}\beta_s + \widetilde{\varepsilon}_{ijs}(W_{ijs}) \tag{1}$$

We want to identify $G_{js,W}$, the distributions of the errors $\widetilde{\varepsilon}_{ijs}(W_{ijs})$, the distribution of r_i, F_r, as well as the coefficients γ_{js} and β_s. Nonparametric identification requires:

1. r_i and $\widetilde{\varepsilon}_{ijs}(W_{ijs})$ are independent and have a median of zero. The variance of r_i is known and finite, and the variances of $\widetilde{\varepsilon}_{ijs}(W_{ijs})$ are finite.
2. $X_{ij1}\beta_1|\gamma_j, W_{ij1}$ is continuous with large support.
3. $X_{ik2}\beta_2|X_{ik1}\beta_1, \gamma_j, \gamma_k, W_{ij1}, W_{ij2}$ is continuous with large support.
4. At least one component of β_1 is assumed equal to the same component in β_2.

The first two conditions are familiar from the standard literature on nonparametric binary choice models. In the second condition, note that identification requires variation in X_{ijs} conditional on regressors W_{ijs} that affect the distribution of errors. The third condition guarantees that there is variation in the regressors conditional on the first-round regressors and judge identity.

Rewriting Equation 1, in each stage an individual is approved if

$$\mathbb{1}[-X_{ijs}\beta_s - \gamma_{js} > \widetilde{\varepsilon}_{ijs}(W_{ijs}) - r_i] = H_{js,W}(-X_{ijs}\beta_s - \gamma_{js}) \tag{2}$$

where $H_{js,W}$ is the distribution of $\eta_{ks} = \widetilde{\varepsilon}_{ijs}(W_{ijs}) - r_i$, the composite error of the refugee-level equality variable r_i and the case-judge idiosyncratic error $\widetilde{\varepsilon}_{ijs}(W_{ijs})$. As in Manski (1975), the assumption of median-zero errors allows nonparametric identification of β_1 and H_{k1} up to scale.

However, the identity of judge j and the regressors W_{ijs} enter the distribution $H_{js,W}$, and thus neither W_{ijs} or the judge effect γ_j can be used for identification. Instead, X_{ij1} traces out the distribution of $H_{js,W}$, which is why Assumption 2 calls for large support conditional on judge assignment and W_{ijs}.

In the second round, the second and third conditions imply that

$$\lim_{X_{ij1}\beta_1\to-\infty}\mathbb{1}[-X_{ik2}\beta_2-\gamma_{k2} > \widetilde{\varepsilon}_{ik2}(W_{ik2})-r_i|-X_{ij1}\beta_1-\gamma_{j1} > \widetilde{\varepsilon}_{ij1}(W_{ij2})-r_i] = H_{k2,W}(-X_{ij2}\beta_2-\gamma_{j2}) \tag{3}$$

so β_2 and H_{k2} are similarly identified to scale. By Assumption 4, this scale is the same. Then, as in Chen et al. (1999), the variances of $\widetilde{\varepsilon}_{ij1}$ and $\widetilde{\varepsilon}_{ij2}$ are identified relative to r_i from the variance of the first and second round residuals and their covariance. Finally, the result of Rao (1971) recovers the distributions $G_{j1,W}$, $G_{j2,W}$, and F_r.

A2 Estimation details

For notational simplicity, I collapse all coefficients and regressors into the distribution of the observational 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.

A2.1 First round approval

$$\begin{aligned} P(r - \varepsilon_1 > 0) &= \int_0^\infty P(r > \tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_1) \\ &= \int_0^{x_m} P(r > \tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_1) + \int_{x_m}^\infty P(r > \tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_1) \end{aligned} \tag{4}$$

The first term in Equation 4 can be shown to be equal to $\Phi[\frac{\ln(x_m)-\mu_1}{\sigma_1}]$. Then,

$$\begin{aligned} \int_{x_m}^\infty P(r > \tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_1) &= \int_{x_m}^\infty \frac{x_m^\alpha}{\tilde{\varepsilon}_1^\alpha} \phi\left(\frac{\ln(\tilde{\varepsilon}_1) - \mu_1}{\sigma_1}\right) \frac{1}{\sigma_1 \tilde{\varepsilon}_1} d\tilde{\varepsilon}_1 \\ &= x_m^\alpha \int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}}^\infty e^{-\alpha(\sigma_1 y + \mu_1)} \phi(y) dy \\ &= x_m^\alpha e^{-\alpha\mu_1} \int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}}^\infty e^{-\alpha\sigma_1 y - \frac{y^2}{2}} dy \\ &= x_m^\alpha e^{-\alpha\mu_1 + \frac{\alpha^2\sigma_1^2}{2}} \int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}}^\infty e^{-\frac{1}{2}(y+\alpha\sigma_1)^2} dy \\ &= x_m^\alpha e^{-\alpha\mu_1 + \frac{\alpha^2\sigma_1^2}{2}} \int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}+\alpha\sigma_1}^\infty e^{-\frac{1}{2}y^2} dy \\ &= x_m^\alpha e^{-\alpha\mu_1 + \frac{\alpha^2\sigma_1^2}{2}} \left[1 - \Phi\left(\frac{\ln(x_m) - \mu_1}{\sigma_1} + \alpha\sigma_1\right)\right] \end{aligned} \tag{5}$$

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

A2.2 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

$$P(r > \varepsilon_2 \cap r > \varepsilon_1) = \int_0^\infty \int_0^\infty P(r > \tilde{\varepsilon}_2 | r > \tilde{\varepsilon}_1) P(r > \tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_2) \tag{6}$$

The terms inside the integrals can be rewritten

$$P(r > \tilde{\varepsilon}_1) = \mathbb{1}[\tilde{\varepsilon}_1 < x_m] + \mathbb{1}[\tilde{\varepsilon}_1 \geq x_m] \frac{x_m^\alpha}{\tilde{\varepsilon}_1^\alpha} \tag{7}$$

and

$$\begin{aligned} P(r > \tilde{\varepsilon}_2 | r > \tilde{\varepsilon}_1) =& \mathbb{1}[\tilde{\varepsilon}_1 < x_m] \left[\mathbb{1}[\tilde{\varepsilon}_2 < x_m] + \mathbb{1}[\tilde{\varepsilon}_2 \geq x_m] \frac{x_m^\alpha}{\tilde{\varepsilon}_2^\alpha} \right] + \\ & \mathbb{1}[\tilde{\varepsilon}_1 \geq x_m] \left[\mathbb{1}[\tilde{\varepsilon}_2 < \tilde{\varepsilon}_1] + \mathbb{1}[\tilde{\varepsilon}_2 \geq \tilde{\varepsilon}_1] \frac{\tilde{\varepsilon}_1^\alpha}{\tilde{\varepsilon}_2^\alpha} \right] \end{aligned} \tag{8}$$

Substituting into Equation 6 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:

$$\int_0^{x_m} \int_0^{x_m} dF(\tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_2) \tag{9}$$

$$\int_{x_m}^\infty \int_0^{x_m} \frac{x_m^\alpha}{\tilde{\varepsilon}_2^\alpha} dF(\tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_2) \tag{10}$$

$$\int_{x_m}^\infty \int_0^{\tilde{\varepsilon}_1} \frac{x_m^\alpha}{\tilde{\varepsilon}_1^\alpha} dF(\tilde{\varepsilon}_2) dF(\tilde{\varepsilon}_1) \tag{11}$$

$$\int_{x_m}^\infty \int_{\tilde{\varepsilon}_1}^\infty \frac{x_m^\alpha}{\tilde{\varepsilon}_2^\alpha} dF(\tilde{\varepsilon}_2) dF(\tilde{\varepsilon}_1) \tag{12}$$

These four equations (9-12) 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

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

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

$$\begin{aligned}\int_{x_m}^{\infty}\int_0^{\tilde{\varepsilon}_1} \frac{x_m^\alpha}{\tilde{\varepsilon}_1^\alpha} dF(\tilde{\varepsilon}_2)dF(\tilde{\varepsilon}_1) &= \int_{x_m}^{\infty}\frac{x_m^\alpha}{\tilde{\varepsilon}_1^\alpha}\Phi\left(\frac{\ln(\tilde{\varepsilon}_1)-\mu_2}{\sigma_2}\right)dF(\tilde{\varepsilon}_1)\\ &= x_m^\alpha\int_{x_m}^{\infty} e^{-\alpha\ln(\tilde{\varepsilon}_1)}\Phi\left(\frac{\ln(\tilde{\varepsilon}_1)-\mu_2}{\sigma_2}\right)\phi\left(\frac{\ln(\tilde{\varepsilon}_1)-\mu_1}{\sigma_1}\right)\frac{1}{\sigma_1\tilde{\varepsilon}_1}d\tilde{\varepsilon}_1\\ &= x_m^\alpha e^{-\alpha\mu_1}\int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}}^{\infty} e^{-\alpha\sigma_1 y}\Phi\left(\frac{\sigma_1 y+\mu_1-\mu_2}{\sigma_2}\right)\phi\left(y\right)dy\\ &= x_m^\alpha e^{-\alpha\mu_1+\frac{\alpha^2\sigma_1^2}{2}}\int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}+\alpha\sigma_1}^{\infty}\Phi\left(\frac{\sigma_1 y-\alpha\sigma_1^2+\mu_1-\mu_2}{\sigma_2}\right)\phi\left(y\right)dy\\ &= x_m^\alpha e^{-\alpha\mu_1+\frac{\alpha^2\sigma_1^2}{2}}\int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}+\alpha\sigma_1}^{\infty}\Phi\left(\frac{y-\alpha\sigma_1+(\mu_1-\mu_2)/\sigma_1}{\sigma_2/\sigma_2}\right)\phi\left(y\right)dy\\ &= x_m^\alpha e^{-\alpha\mu_1+\frac{\alpha^2\sigma_1^2}{2}} BvN\left(\frac{(\mu_1-\mu_2)/\sigma_1-\alpha\sigma_1}{\sqrt{\sigma_2^2/\sigma_1^2+1}}, -\frac{\ln(x_m)-\mu_1}{\sigma_1}-\alpha\sigma_1, \frac{1}{\sqrt{\sigma_2^2/\sigma_1^2+1}},\right)\end{aligned}$$

$$
\begin{aligned}
\int_{x_m}^{\infty}\int_{\tilde{\varepsilon}_1}^{\infty}\frac{x_m^{\alpha}}{\tilde{\varepsilon}_2^{\alpha}}dF_1dF_2 &= \int_{x_m}^{\infty} x_m^{\alpha}e^{-\alpha\mu_2+\frac{\alpha^2\sigma_2^2}{2}}\left[1-\Phi\left(\frac{\ln(\tilde{\varepsilon}_1)-\mu_2}{\sigma_2}+\alpha\sigma_1\right)\right]dF(\tilde{\varepsilon}_1) \\
&= \tilde{B}\left\{\left[1-\Phi\left(\frac{\ln(x_m)-\mu_1}{\sigma_1}\right)\right]-\int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}}^{\infty}\Phi\left(\frac{y+(\mu_1-\mu_2+\alpha\sigma_2^2)/\sigma_1}{\sigma_2/\sigma_1}\right)\phi(y)dy\right\} \\
&= \tilde{B}\left\{\left[1-\Phi\left(\frac{\ln(x_m)-\mu_1}{\sigma_1}\right)\right]-BvN(\frac{(\mu_1-\mu_2+\alpha\sigma_2^2)/\sigma_1}{\sqrt{\sigma_2^2/\sigma_1^2+1}},-\frac{\ln(x_m)-\mu_1}{\sigma_1},\frac{1}{\sqrt{\sigma_2^2/\sigma_1^2+1}})\right\}
\end{aligned}
$$

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

A2.3 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 12 in detail; the same method works for all the joint first- and second-round probabilities.

$$
\begin{aligned}
&\int_{x_m}^{\infty}\int_{\tilde{\varepsilon}_1}^{\infty}\frac{x_m^{\alpha}}{\tilde{\varepsilon}_2^{\alpha}}dF_2dF_1 \\
=&\int_{x_m}^{\infty}\int_{\tilde{\varepsilon}_1}^{\infty}\frac{x_m^{\alpha}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}\varepsilon_1\varepsilon_2}e^{-\alpha\ln(\varepsilon_2)-\frac{1}{2(1-\rho^2)}\left((\frac{\ln(\varepsilon_1)-\mu_1}{\sigma_1})^2-2\rho(\frac{\ln(\varepsilon_1)-\mu_1}{\sigma_1})(\frac{\ln(\varepsilon_2)-\mu_2}{\sigma_2})+(\frac{\ln(\varepsilon_2)-\mu_2}{\sigma_2})^2\right)}d\varepsilon_2d\varepsilon_1 \\
=&x_m^{\alpha}e^{-\alpha\mu_2}\int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}}^{\infty}\int_{\frac{\sigma_1y+\mu_1-\mu_2}{\sigma_2}}^{\infty}\frac{1}{2\pi\sqrt{1-\rho^2}}e^{\frac{-1}{2(1-p^2)}\left(2\alpha\sigma_2(1-p^2)x+y^2-2\rho yx+x^2\right)}dxdy
\end{aligned}
$$

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

$$
x_m^{\alpha}e^{-\alpha\mu_2+\frac{\alpha^2\sigma_2^2}{2}}\int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}}^{\infty}\left(1-\Phi(\frac{\frac{\sigma_1y+\mu_1-\mu_2}{\sigma_2}-(\rho y-\alpha\sigma_2(1-\rho^2))}{\sqrt{1-\rho^2}})\right)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y+\alpha\sigma_2\rho)^2}dy
$$

Rearrange the term in the normal:

$$
\frac{\frac{\sigma_1y+\mu_1-\mu_2}{\sigma_2}-(\rho y-\alpha\sigma_2(1-\rho^2))}{\sqrt{1-\rho^2}} = \frac{y+(\mu_1-\mu_2+\alpha\sigma_2^2(1-\rho^2))/(\sigma_1-\rho\sigma_2)}{\sigma_2\sqrt{1-\rho^2}/(\sigma_1-\rho\sigma_2)}
$$

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.

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

$$
\begin{aligned}
\widetilde{B} &= x_m^{\alpha} e^{-\alpha\mu_2+\frac{\alpha^2\sigma_2^2}{2}} \\
b &= \alpha\sigma_2\rho-(\mu_1-\mu_2+\alpha\sigma_2^2(1-\rho^2))/(\sigma_1-\rho\sigma_2) \\
a &= \sigma_2\sqrt{1-\rho^2}/(\sigma_1-\rho\sigma_2)
\end{aligned}
$$

A3 MTE of first-round approval on second-round approval

A natural question to ask is how likely individuals approved in the first round are to be ultimately successful in the second round. The Federal Court's own standard is that individuals should be granted leave in the first round if they can make an "arguable case" in the second round. A simple way to quantify this is to estimate the MTE of first-round approval on second-round approval. In the notation of Heckman and Vytlacil (2005), this is

$$\Delta^{MTE}(u_D) = E[Y_1 - Y_0 | U_D = u_d] \tag{13}$$

where Y is second-round approval. In this context Y_0 is mechanically equal to zero (you cannot be approved in the second round if you aren't approved in the first). Treatment (or first-round approval) is determined by

$$D^* = P(Z) \geq U_D \tag{14}$$

where U_D is normalized to be unit uniform and $P(Z)$ is the probability of treatment given assignment to the instrument Z. I use the first-round judge assignment as the instrument Z. As seen in Figure A1, the support of the instrument ranges from 0.03 to 0.28, with a large gap before a point mass at 0.70. I estimate the MTE using all observations, and using only the main mass. Nonetheless, the results are only identified by functional form for points larger than 0.3, and should be treated with caution.

As Figure A1 shows, there is not very much variation in MTE over the range of first-round judges; from the 3rd to the 28th percentile of the distribution of refugee quality, the approval probability drops from 46% to 35%, though this result is somewhat sensitive to the outlier judge. In other words, most judges' marginal rejections would have at least a one-third chance of approval in the second round.

A4 Survey questions

As I discuss in Section 4.8, I fielded a survey of lawyers who had appeared in front of the Federal Court justices in my sample. The goal of the survey was to generate expert measures of the same parameters that are identified by my structural model.

From the court records, I located the names of 931 lawyers who had appeared in front of one of the judges in my sample. I was able to find online contact information for 551 of them.[1] In April 2017, I contacted the lawyers and requested that they fill out an online survey on their experience with Federal Court judges. After one reminder email, 64 lawyers responded for an overall response rate of 14%.[2] Table A4 compares responders to non-responders and lawyers for whom I couldn't find contact information. The main differences are that responders are more successful, with a first-round approval rate of 27% versus 19% for non-responders (the contacted sample is mostly lawyers for the claimants; government lawyers were included in the sample but their names are recorded much less frequently in the court documents). Respondents are slightly younger, with their first recorded case coming about one year later.

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

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

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

2. On a scale from 1 to 5, how would you rate the listed judges in terms of **consistency?** Are their decisions predictable compared to other judges with similar grant rates? Do they decide cases on similar grounds as other justices? Can you predict what grounds the case will be decided on?

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

[2] This response rate compares favorably to telephone political polls, where response rates are below 10% (Pew, 2012). However, it is significantly lower than the 20% response rate for an email poll conducted by Card et al. (2012) surveying UC Berkeley staff about job satisfaction. The difference in response rates is likely due to declining survey rates over time (Card et. al surveyed in 2008), a pecuniary incentive, and that they had the advantage of being able to present themselves as in-group members (other University of California employees).

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

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

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

I expected that the first question would be related to the judge-specific threshold γ, and the second question with the variance of the observational error σ_j. By design, the second question encompasses the two distinct aspects of σ_j detailed in Section 2.3.1. First, asking about predictability concerns test-retest reliability — will the judge understand the merits of the case? On the other hand, asking whether the judge decides cases on similar grounds as other judges is trying to unearth information about how judges consistently value different aspects of the case (inter-rater reliability), such as the relative weight they place on procedural versus substantive merits.

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

The main results are in Table 6, where I include only the first two questions. I discuss these in Section 4.8. The final question of the survey, which asked about how accurate the judge is, I did not discuss in the main text. This question does not have as clear an interpretation as the other two. There is no direct mapping of accuracy into the model, since accuracy implies a normative judgement about the correct outcome of the case. Reported accuracy is correlated with favorability and consistency, but more strongly with the former ($\rho = 0.7$ versus 0.46). Anecdotally, many of the lawyers that I corresponded with about the survey were involved in refugee-rights non-profits, so it is likely that they believe the claimants should win more cases than they currently do. Table A5 adds accuracy to the regression of model coefficients on survey responses; with no other regressors higher accuracy predicts lower second-stage thresholds γ_{j2}, but this disappears when favorability and consistency are added. The relationship between favorability and γ_2, and consistency and σ_1 is almost unchanged.

A5 Ramifications for judge-assignment IVs

Exploiting random judge assignment is an increasingly popular identification strategy (Aizer and Doyle, 2013; Dahl et al., 2013). The monotonicity condition in this context is simple: it requires that for any two judges, any individual convicted by the more lenient judge must also be convicted by the less lenient judge.[3] Mueller-Smith (2014) discusses how this can be violated when the researcher does not separately estimate judge severity by type of crime. If a judge is harsh for (say) drug crimes but lenient for violent crimes, on average they would be considered a medium-severity judge. But exposure to this judge versus one that uniformly sentences defendants for an average sentence would be a negative shock for a drug-crime defendant and a positive shock for a violent-crime defendant, generating defiers. Mueller-Smith shows how this problem can be circumvented when the econometrician has access to a rich set of covariates. The strategy is to use a LASSO first stage to select the instruments (in his case, judge and prosecutor effects interacted with defendant characteristics and crime type) with the best predictive power, ensuring that inconsistency associated with the observables is not contributing to violations of monotonicity. Other researchers have approximated this approach with simple interactions between judge assignment and crime type (Norris and Pecenco, 2017). Bhuller et al. (2016) show that their judge-assignment instrument (judge-mean incarceration rate on all cases) also predicts incarceration for subsets of the sample defined by defendant and charge characteristics. In most situations, these methods are likely to assuage first-order concerns about monotonicity violations.

As I show in the main text, in my context there are high levels of inconsistency (and corresponding violations of monotonicity). This is not necessarily an indictment of previous research that relies on examiner-assignment instruments. As I discuss above, it is possible to partially test for monotonicity and construct instruments that are robust against most sorts of violations. Furthermore, Federal Court refugee appeals are very different than criminal courts or the SSDI system, and it is possible that these decisions are more susceptible to inconsistency for two reasons. First, appeal cases are almost by definition more marginal. About 60% of initially-rejected claimants appeal to the Federal Court, meaning that the cases on the docket are relatively difficult. Second, as I discuss in Section 3.2, there is no written precedent for first stage decisions. This makes it hard for judges to learn how other judges have acted in a similar situation. I show suggestive evidence in the main text that this manifests itself in a wider cross-judge distribution of first-round thresholds γ_{j1} (relative to the second round); it is likely that it also results in lower consistency.

With the above caveats, it is worthwhile to examine how inconsistency affects estimates of the marginal treatment effect arising from my data. The presence of inconsistency breaks the relationship between the approval rate of the judge and the identity of the judge's marginal claimants

[3] De Chaisemartin (2017) shows that even under violations of monotonicity it may be possible to identify a less-interpretable LATE for a subset of compliers.

— there is instead a distribution of marginal claimants. Estimates of the marginal treatment effect under inconsistency thus average together individuals with different case strengths (Klein, 2010). I quantify the implications of inconsistency in two ways. First, I use my empirical setting to directly calculate how the estimated MTE of first-round approval on second-round approval changes with consistency. This approach is close to the data, and demonstrates that MTE estimates under inconsistency may be flatter than the underlying MTE. Second, I use the method of Klein (2010) to calculate the estimated bias under different theoretical MTEs that might be encountered in other contexts.

A5.1 MTE bias estimated from data

There is not a single way to quantify the effect of inconsistency on the estimated MTE. There are many potential non-degenerate joint distributions of first-round judge errors $\widetilde{\varepsilon}_{ij1}$that do *not* generate violations of monotonicity — for example, if all judges had the same error $\widetilde{\varepsilon}_{ij1}$ for each claimant. This is important because although the estimated parameters guarantee violations of monotonicity (recall from Section 4.3 that I bound the share of cases judge pairs disagree on above zero), different assumptions on the cross-judge joint distribution of $\widetilde{\varepsilon}_{ij1}$ allow for different counterfactuals that satisfy the monotonicity assumption. For simplicity I choose the most straightforward alternative: judges are perfectly consistent ($\sigma_{j1} = 0$ for all judges), guaranteeing montonicity is satisfied. I adjust thresholds γ_{j1} to keep the approval rate the same for all judges, then calculate the MTE under both the baseline coefficients and the counterfactual. Figure A3 shows that as consistency declines, the estimated MTE becomes shallower. This reflects how inconsistency generates a distribution of marginal claimants for each judge, each with a different treatment effect. Averaging over the distribution of marginal claimants for each judge reduces the cross-judge variation in the estimated MTE.

Interestingly, the simulated IV coefficient of second-stage approval on first-stage approval instrumented by judge-mean approval barely changes, from 0.251 to 0.265. This suggests that inconsistency in judge-assignment designs may more strongly affect the MTE than the IV estimate.

A5.2 MTE bias for hypothetical MTEs

A second method to quantify the effect of inconsistency on estimated MTE comes from Klein (2010). His approach is to take a second-order expansion of the MTE estimate around a baseline of perfect consistency, then study how estimates change. I use the judge-specific estimates from the first round for the selection stage, and then estimate the bias under three different MTEs. I follow Klein and use the same functional form for the MTE as in his empirical example,

$$m(v) = 5 + 1.5 * (1 - v)^{\rho}$$

where I pick $\rho = \{0.5, 1, 1.5\}$. I assume that the true MTE is with respect to the refugee factor; $v = 1 - F_r(r_i)$. Panel A of Figure A4 plot these MTEs.

Heuristically, judicial errors bias the MTE estimate by replacing a point estimate with an estimate of the local average MTE. Klein shows that this error can be approximated by

$$\frac{1}{2}\sigma_p^2 \frac{\partial^2 m(p)}{\partial p^2} + \frac{1}{2}\frac{\partial \sigma_p^2}{\partial p}\frac{\partial m(p)}{\partial p} \tag{15}$$

where $m(p)$ is the marginal treatment effect with respect to the instrument-induced participation probability p and σ_p^2 represents stochastic variation in whether an individual will be induced into treatment by a particular value of the instrument. It is similar to my measure of inconsistency, σ_{js}. MTE bias results from both curvature of the MTE interacted with the size of inconsistency, and cross-judge *changes* in inconsistency interacted with the slope of the MTE.

Panels B-D of Figure A4 plot estimates of the MTE bias at each point in the support of the instrument. As expected, the bias is worse in areas of the MTE with higher curvature, and worse for more-steeply sloped MTEs. The IV biases are 10, 19, and 26% for ρ of 0.5, 1, and 1.5. In this context, bias results more from the slope of the MTE (the second term in Equation 15) than from curvature (first term).

Interestingly, and in contrast to the application-specific method in Section A5.1, the Klein method predicts negative bias over the support of the instrument. The cause of this difference is subtle. In Section A5.1, I kept overall approval rates the same while adjusting consistency σ_{j1}. Klein does not make this restriction, and so the results of Figure A4 partially reflect the addition of a large number of treated individuals. Because the distribution of underlying case strength r_i is right-tailed, the additional marginally treated claimants have disproportionately weak cases. This biases the estimated MTE downwards at all points.

The overall message of this section is one of cautious optimism. My context may be a worst-case scenario for monotonicity: the cases are relatively difficult, there are no case type controls to interact with judge effects, and the lack of precedent likely contributes to inconsistency. Despite this, worst-case estimates of the bias in LATE estimates are only about 25%. With lower levels of inconsistency or a relatively flat MTE, the size of the IV bias is likely to be small.

A6 Model parameters without additional regressors

The baseline model uses dummies for a late-week decision and whether the second-round hearing was made over lunch to aid in identification. In this section I present estimates from a model identified without regressors, as well as the main results. Identification now leans more strongly on functional form, though judge randomization still identifies relative consistency for judges with similar approval rates. Figure A5 presents the coefficients. The raw coefficients are similar to the baseline model but less precisely estimated. This reassuringly suggests that the results are driven mainly by the judge randomization (and to some degree the functional forms), but that the use of regressors additionally improves precision.

In Section 4.6 I present evidence that judicial inconsistency is due more to idiosyncratic observational errors than permanent ideological differences between judges in statutory interpretation. One fear with this approach is that errors arising from a late-week and noon-time decisions may be precisely idiosyncratic errors rather than ideological 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 4, but uses the no-regressor model probabilities. The results are comparable to the baseline specification: the model predicts second-stage approval well, but conditional on the model probabilities there is little additional predictive power from knowing the exact judge pairs. The distribution of the EB means of the judge pairs is very similar — in my preferred, rightmost specification, the standard deviation of the judge pair effects is 0.003 in the both models — and the F-test similarly does not reject that the judge pair effects are jointly zero.

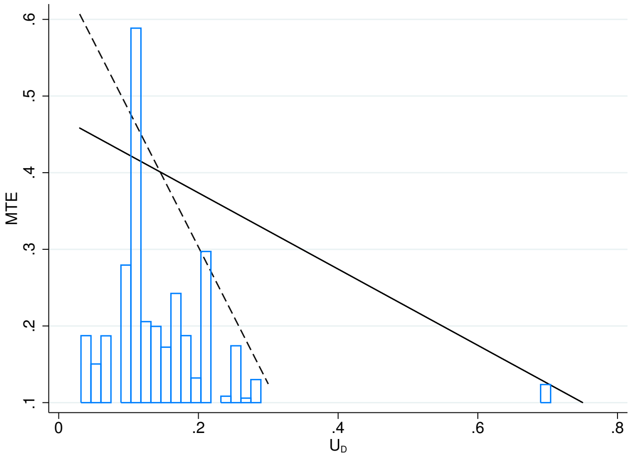
In Table A9, I test the effect of experience and workload on inconsistency. Similarly to Table 5, I find that judges become dramatically more consistent after one year of experience, but continue to make gains through at least the first ten years on the job. Higher caseloads decrease consistency (Columns 3 and 4), though only for judges with fewer than 6 years of experience (Column 5).

Table A10 contains estimates of the effect of judicial selection reform on judge consistency. As I describe in Section 3.3, changes to the laws governing judicial selections made it much more difficult for governments to grant judgeships to unqualified party supporters after 1988. In Section 4.9 I show that this reduced baseline estimates of consistency by approximately 75%. In the model estimated without regressors, the results are large but not quite so dramatic — in the baseline specification inconsistency declines by 0.7 from a pre-reform mean of 1.7.

Finally, Table A11 mirrors the results of the baseline model: estimates of judge thresholds γ_{js} are negatively correlated with survey measures of judge favorability to claimants, and model-estimated inconsistency σ_{j1} is negatively correlated with surveyed consistency. Again, this suggests that the correlation between model and survey results are not driven by the use of regressors in identification.

A7 Appendix Figures

Figure A1: MTE of second-round approval on first-round approval judge



The black line represents the MTE of first-round approval on second-round approval (implicitly, no one who is rejected in the first round is approved in the second round). Estimation is from regressing a second-round approval on a second-order polynomial in the judge-level first-round approval means, then taking the analytic derivative. First-round approval is instrumented by judge assignment. The distribution of judge-mean approval rates is displayed as a histogram. The dashed line is the MTE estimated without the outlier point.

Figure A2: Scatter plot of first- and second-round consistency σ_{js}

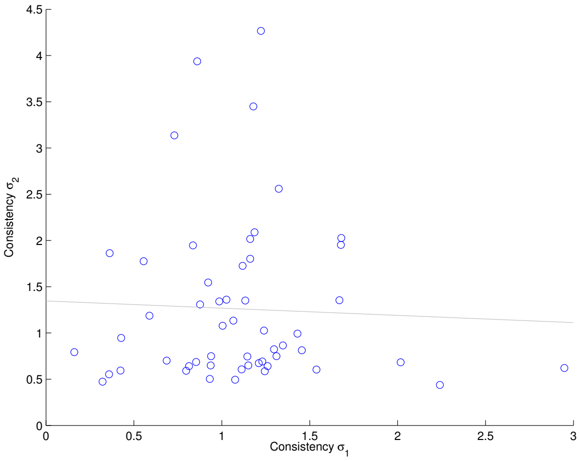


Figure A3: Estimated MTE at baseline and under consistency

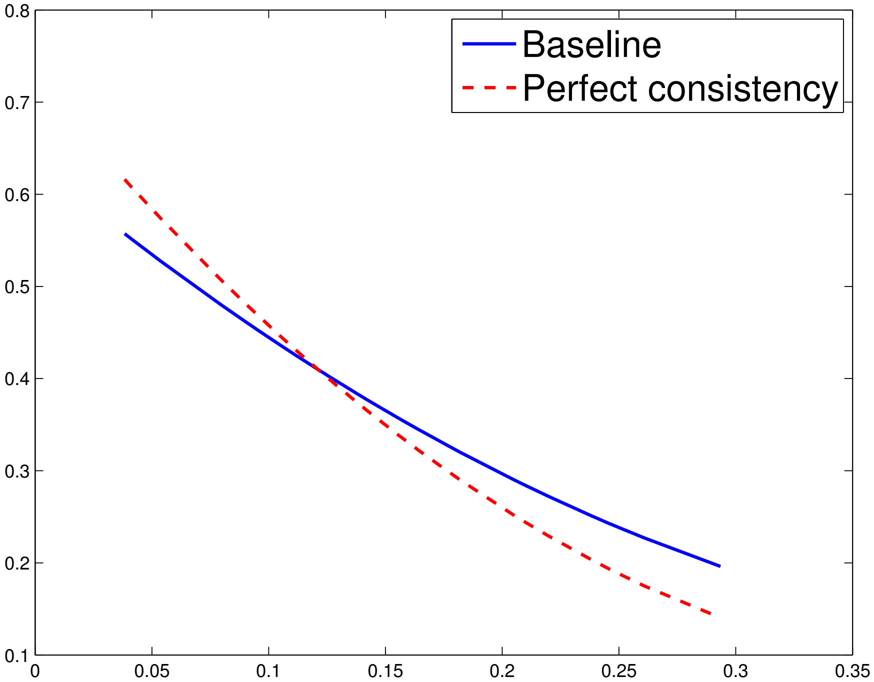
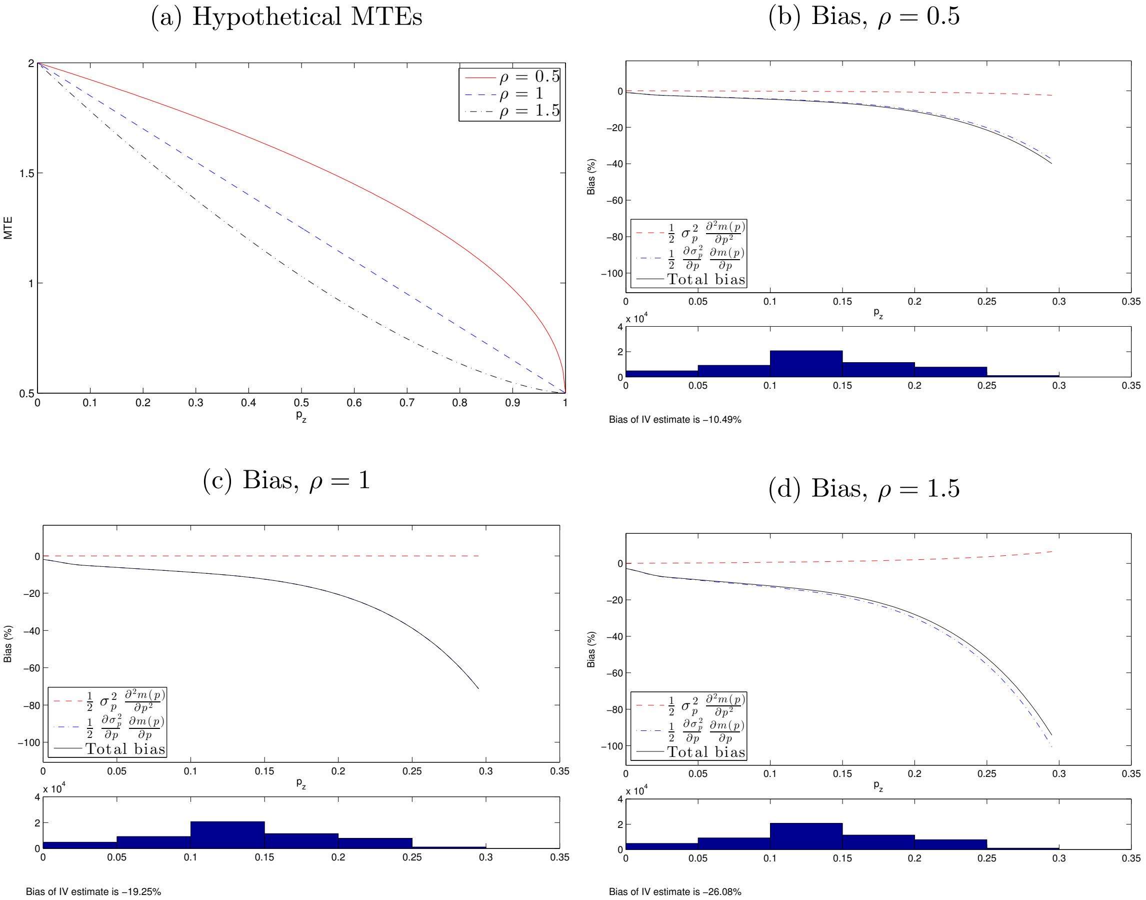


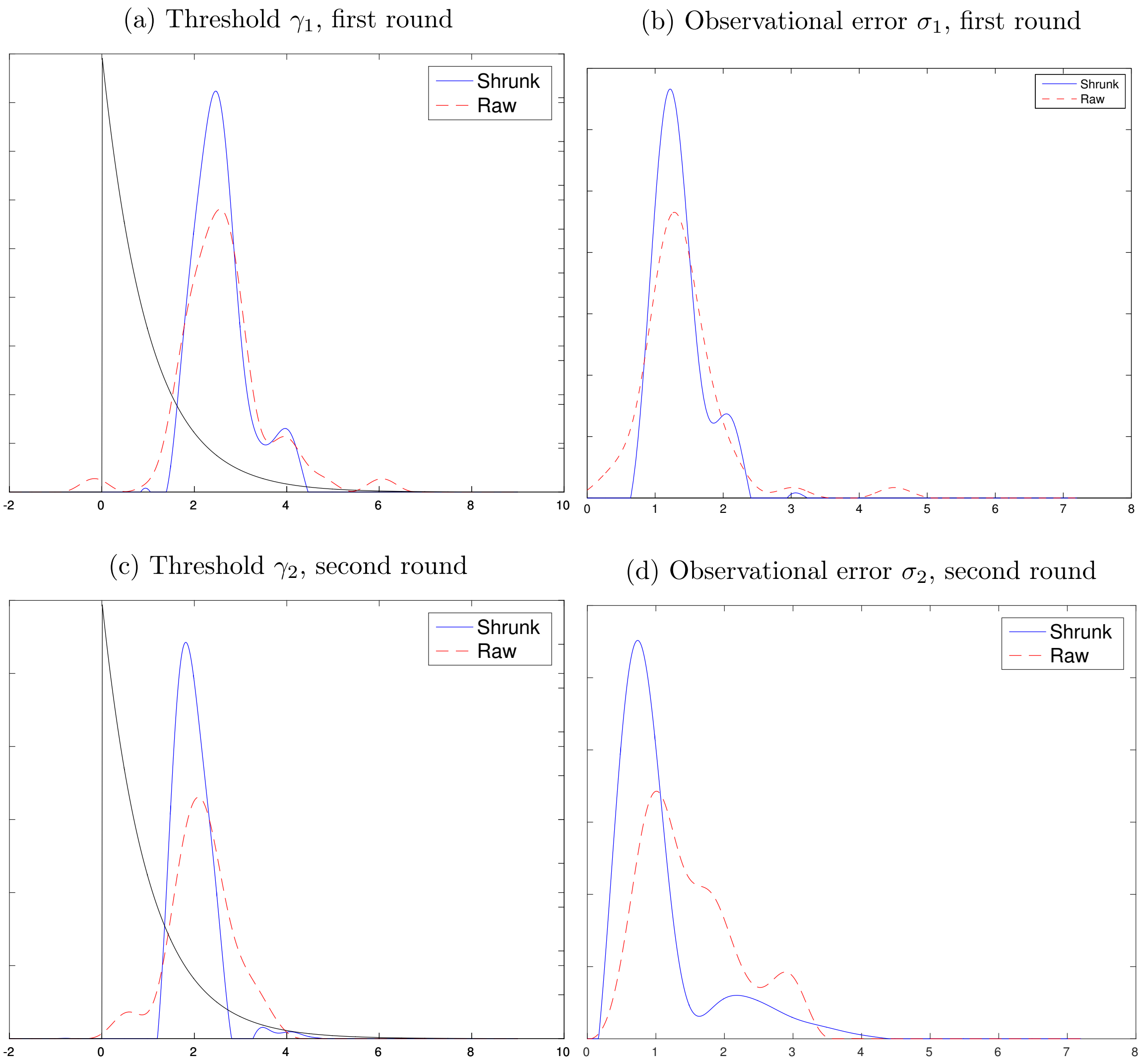
Figure demonstrates how inconsistency affects estimate of MTE. I plot the baseline MTE in solid blue, then use model estimates to construct an estimate of the MTE if all judges were perfectly consistent (ie, $\sigma_{j1} = 0$) but had the same average approval rate.

Figure A4: Bias for different MTEs



Panel A plots the marginal treatment effects $m(v) = 5 + 1.5 * (1 - v)^{\rho}$ for $\rho = \{0.5, 1, 1.5\}$. Panels B, C and D plot bias over the support of the main mass of the instrument for each MTE.

Figure A5: Distribution of judge coefficients, model identified without regressors



(a) Threshold γ_1, first round

(b) Observational error σ_1, first round

(c) Threshold γ_2, second round

(d) Observational error σ_2, second round

Figure displays coefficients for decision model $\mathbb{1}[r_i > \gamma_{js} + \widetilde{\varepsilon}_{ijs}]$, $\widetilde{\varepsilon}_{ijs} \sim \mathcal{N}(0, \sigma^2_{js})$. All models allow the parameters of the Pareto distribution of r_i to vary flexibly by office of origination as well as after 2002 relative to before 2002, when there were legislative changes that may have impacted the distribution of case quality. Each panel contains the density of the raw and shrunken estimates of the judge thresholds γ_1 and γ_2, and judge inconsistency σ_1 and σ_2. Black line is density of case quality r_i. Shrunken estimates recovered via deconvolution of estimates accounting for coefficient-specific standard errors (Delaigle and Meister, 2008).

A8 Appendix Tables

Table A1: Judge 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: Randomization using name-imputed continent of origin

	Male	Africa	Asia	South America	Predicted approval	1st-round mean approval
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: First round judges						
First-round mean approva ratel	0.003	-0.086***	0.018	0.080	-0.001	
	(0.018)	(0.031)	(0.027)	(0.065)	(0.001)	
F-stat	0.87	2.90	3.04	6.65	3.51	
Prob	0.75	0.00	0.00	0.00	0.00	
Observations	58,604	58,604	58,604	58,604	58,604	
Panel B: Second round judges						
Second-round mean approval rate	-0.025	-0.005	-0.012	0.043	-0.001	-0.025
	(0.033)	(0.031)	(0.035)	(0.039)	(0.002)	(0.016)
F-stat	1.07	1.80	1.17	1.62	1.57	4.33
Prob	0.33	0.00	0.17	0.00	0.00	0.00
Observations	8,446	8,446	8,446	8,446	8,446	8,446

Gender and continent of origin predicted from claimant name. IRB mean approval is the approval rate of the IRB Member who initially denied refugee status to the claimant. Predicted approval comes from a regression of approval on gender, continent of origin and IRB Member approval rate. F-stats come from separate regression of outcome on judge fixed effects. All regressions include office X pre-2002 fixed effects to account for cross-office differences in case strength and changes in government policy in 2002. Standard errors clustered at the judge level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: Testing effect of regressors on distribution of judge errors

	(1)	(2)	(3)
Coefficients β affecting judge threshold γ_1			
End-of-week decision	0.057***	0.087***	0.051**
	(0.004)	(0.029)	(0.021)
Hearing schedule over lunch	0.411***	0.381**	0.510***
	(0.077)	(0.193)	(0.165)
Coefficients ψ affecting judge inconsistency σ_1			
End-of-week decision		0.040	
		(0.058)	
Hearing schedule over lunch			0.371
			(1.287)
SD of γ_1	0.836	0.833	0.840
SD of σ_1	0.485	0.467	0.475

Reports coefficients for choice model $\mathbb{1}[r_i > \gamma_{js} + X_{ijs}\beta_s + \widetilde{\varepsilon}_{ijs}(W_{ijs})]$, $\widetilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\widetilde{\sigma}_{js} + W_{ijs}\psi_s)$. 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. Standard errors clustered at the level of the first stage judge. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Lawyer characteristics, survey respondents vs lawyer population

	Respondents	NR/NC	Difference
Success rate (first round)	0.27	0.19	0.078***
	[0.22]	[0.21]	(0.027)
Success rate (second round)	0.13	0.08	0.049***
	[0.16]	[0.15]	(0.019)
First case (year)	2002.55	2001.37	1.179*
	[5.36]	[5.39]	(0.698)
Number of cases (total)	141.77	101.62	40.149
	[225.93]	[221.69]	(28.752)
Male (=1)	0.67	0.60	0.067
	[0.47]	[0.48]	(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 A5: Model coefficients on survey responses including accuracy response

	(1)	(2)	(3)	(4)
Panel A: Threshold γ_1 (mean=2.49, SD=1.05)				
Accuracy, SD	-0.130	-0.136*	0.111	0.128
	(0.096)	(0.076)	(0.172)	(0.112)
Favorability, SD			-0.221	-0.315*
			(0.260)	(0.184)
Consistency, SD			-0.153*	-0.075
			(0.087)	(0.069)
Respondent FE	No	Yes	No	Yes
Observations	174	174	174	174
Panel B: Threshold γ_2 (mean=2.16, SD=1.46)				
Accuracy, SD	-0.195***	-0.308***	-0.016	-0.167
	(0.059)	(0.079)	(0.078)	(0.104)
Favorability, SD			-0.225***	-0.309***
			(0.083)	(0.107)
Consistency, SD			-0.039	0.133*
			(0.055)	(0.072)
Respondent FE	No	Yes	No	Yes
Observations	174	174	174	174
Panel C: Consistency σ_1 (mean=2.06, SD=2.15)				
Accuracy, SD	0.029	0.092	0.011	0.149
	(0.073)	(0.083)	(0.099)	(0.113)
Favorability, SD			0.123	0.052
			(0.124)	(0.111)
Consistency, SD			-0.126*	-0.168**
			(0.067)	(0.078)
Respondent FE	No	Yes	No	Yes
Observations	174	174	174	174

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

Table A6: Approval rate for judges before and after reform

	Approval rate			Approval, year residualized		
	(1)	(2)	(3)	(4)	(5)	(6)
After 1988 reform (=1)	0.0199	0.0490	0.0485	0.0204	0.0656	0.0656
	(0.0272)	(0.0546)	(0.0559)	(0.0260)	(0.0541)	(0.0555)
Liberal appointee (=1)			-0.00621			0.000522
			(0.0211)			(0.0213)
Year appointed	No	Yes	Yes	No	Yes	Yes
N judges	53	53	53	53	53	53

Robust standard errors in parentheses and clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A7: Second-round inconsistency for judges before and after reform

	Baseline			Year control in σ_2		
	(1)	(2)	(3)	(4)	(5)	(6)
After 1988 reform (=1)	-0.0448	0.425	0.297	-1.432***	-0.994	-0.450
	(0.290)	(0.377)	(0.391)	(0.311)	(0.816)	(0.842)
Liberal appointee (=1)			0.0973			-0.651
			(0.175)			(0.397)
Male judge (=1)			-0.453**			-0.817**
			(0.198)			(0.358)
Year appointed	No	Yes	Yes	No	Yes	Yes
Dependent mean	0.92	0.92	0.92	1.95	1.95	1.95
Pre-reform mean	0.96	0.96	0.96	3.31	3.31	3.31
N judges	53	53	53	53	53	53

Estimated with Hanushek (1974) correction for estimated dependent variable. Dependent variable is consistency σ_{j2}, which is estimated from decision model $\mathbb{1}[r_i > X_{ijs}\beta_s + \gamma_{js} + \widetilde{\varepsilon}_{ijs}(W_{ijs})]$, $\widetilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\widetilde{\sigma}_{js} + W_{ijs}\psi_s)$. 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 A8: Second-round outcome on model approval probability and judge-pair FEs

	Judge-pair round FEs		Judge-pair FEs	
	(1)	(2)	(3)	(4)
Model approval probability	0.949^{***}	0.947^{***}	0.976^{***}	0.975^{***}
	(0.166)	(0.166)	(0.0478)	(0.0478)
Model controls	No	Yes	No	Yes
Mean approval	0.44	0.44	0.44	0.44
F-stat for judge pairs	1.02	1.02	0.98	0.99
P-value	0.318	0.314	0.641	0.610
BS p-value	0.581	0.627	0.556	0.544
SD of judge-pair EB means	0.006	0.006	0.004	0.003
Observations	8,196	8,196	8,196	8,196

Regresses second-round approval on model-predicted likelihood of approval and judge-pair fixed effects. In contrast to Table 4, 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 consistency by experience and workload

	(1)	(2)	(3)	(4)	(5)
Coefficients ψ affecting judge inconsistency σ_1					
Experience > 1 year	-0.891	-0.797^{**}		-0.557^{**}	-0.462^{**}
	(0.993)	(0.365)		(0.267)	(0.203)
Experience > 5 years		-0.351^{***}		-0.049	-0.054
		(0.104)		(0.561)	(0.380)
Experience > 10 years		-0.530		-0.447^{***}	-0.501^{***}
		(0.524)		(0.170)	(0.012)
Log caseload			0.204^{***}	0.139^{***}	
			(0.020)	(0.038)	
Log caseload ($\leq$ 5 yrs exp)					0.179^{***}
					(0.025)
Log caseload (> 5 yrs exp)					0.056
					(0.146)
Second-round experience control	Yes	Yes	No	Yes	Yes

Reports coefficients for decision model $\mathbb{1}[r_i > \gamma_{js} + X_{ijs}\beta_s + \widetilde{\varepsilon}_{ijs}(W_{ijs})]$, $\widetilde{\varepsilon}_{ijs} \sim \mathcal{N}(0, \sigma^2_{js})$, $\sigma_{js}(W_{ijs}) = e^{\widetilde{\sigma}_{js}}$. In contrast to the baseline model, the reported models do not use timing regressors for identification. 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. Standard errors clustered at the level of the first stage judge. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A10: Inconsistency before and after judge selection reform

	Baseline			Experience control in σ_1		
	(1)	(2)	(3)	(4)	(5)	(6)
After reform (=1)	-0.154	-0.703**	-0.820***	-1.096***	-0.745**	-0.698*
	(0.150)	(0.287)	(0.283)	(0.307)	(0.339)	(0.376)
Liberal appointee (=1)			0.0549			-0.0788
			(0.135)			(0.0797)
Male judge (=1)			-0.409**			0.00181
			(0.168)			(0.119)
Year appointed	No	Yes	Yes	No	Yes	Yes
Pre-reform mean	1.49	1.49	1.49	1.71	1.71	1.71
N judges	53	53	53	53	53	53

Estimated with Hanushek (1974) correction for estimated dependent variable. Dependent variable is consistency σ_{j1}, which is estimated from decision model $\mathbb{1}[r_i > \gamma_{js} + X_{ijs}\beta_s + \widetilde{\varepsilon}_{ijs}(W_{ijs})]$, $\widetilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\widetilde{\sigma}_{js} + W_{ijs}\psi_s)$. In contrast to the baseline model, the reported models do not use timing regressors for identification. 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: Model coefficients on survey responses

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: γ_1 *(mean=2.65, SD=1.11)*						
Favorability, SD	-0.314	-0.328**			-0.267	-0.321*
	(0.265)	(0.163)			(0.260)	(0.181)
Consistency, SD			-0.247**	-0.135**	-0.180***	-0.022
			(0.100)	(0.060)	(0.057)	(0.083)
Respondent FE	No	Yes	No	Yes	No	Yes
Observations	182	182	182	182	182	182
Panel B: γ_2 *(mean=2.24, SD=1.22)*						
Favorability, SD	-0.315***	-0.435***			-0.329***	-0.487***
	(0.117)	(0.151)			(0.118)	(0.156)
Consistency, SD			-0.046	-0.071	0.045	0.123*
			(0.068)	(0.087)	(0.064)	(0.073)
Respondent FE	No	Yes	No	Yes	No	Yes
Observations	182	182	182	182	182	182
Panel C: σ_1 *(mean=2.28, SD=2.19)*						
Favorability, SD	0.130	0.065			0.187**	0.138
	(0.083)	(0.105)			(0.077)	(0.105)
Consistency, SD			-0.110	-0.126**	-0.167**	-0.185***
			(0.070)	(0.059)	(0.065)	(0.063)
Respondent FE	No	Yes	No	Yes	No	Yes
Observations	182	182	182	182	182	182

Reports linear regressions of model coefficients on survey responses, estimated with Hanushek (1974) correction for estimated dependent variable. Decision model is $\mathbb{1}[r_i > \gamma_{js} + \widetilde{\varepsilon}_{ijs}]$, $\widetilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\widetilde{\sigma}_{js} + W_{ijs}\psi_s)$. In contrast to the baseline model, the reported model does not use timing regressors for identification. The parameters of the Pareto distribution of r_i vary flexibly by office of origination as well as after 2002 relative to before 2002, when there were legislative changes that may have impacted the distribution of case quality. I adjust for judge experience in β_s and ψ_s. Model standard errors clustered at the level of the first stage judge, linear standard errors at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.