Judicial Errors: Evidence from Refugee Appeals*

Samuel Norris†

October 3, 2017

PRELIMINARY AND INCOMPLETE - PLEASE DO NOT DISTRIBUTE

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, and measuring second-round approval rates as a function of first-round judge severity. Structural estimates of judicial inconsistency for a sample of Canadian refugee appeals are highly correlated with lawyers' surveyed opinions on judge ability. Overall levels of judicial consistency are low relative to the distribution of case strength; judges who approve the same share of claimants disagree on 18% of the approved claimants. However, judges become much more consistent with 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 political 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 14,400 cases would be successful over 1995-2012, versus 3,700 under current policy.

[Most recent version here]

*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, Lori Hausegger, 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. I am grateful to Catherine Dauvergne, Sean Rehaag and the many anonymous judges and law clerks who offered valuable insight into the institutional details of the Immigration and Refugee Board and the Federal Court. Paul Longley generous 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 non-traffic cases 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) make the more specific point that inefficient courts increase transaction costs and reduce 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 et al., 2014), inter-communal violence unrelated to the crime (Shayo and Zussman, 2010), the previous decision (Chen et al., 2016), the timing of the hearing relative to lunch (Danziger et al., 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 reliabiity of individual judges. Do judges differ only in *leniency*, the share of defendants they incarcerate? Or do judges with the same overall incarceration rate also vary in *consistency*, the ability to select the defendants their colleagues would agree are guilty? The tools I develop in this paper allow me to separately identify these measures, which facilitate evaluations of the efficiency of the judicial system (are guilty defendants likely to be incarcerated?) and allow the researcher to evaluate the success of reforms. Consistency is closely related to the montonicity assumption of judge-assignment IV designs, and my results shed light on potential biases in this increasingly-common identification strategy (Dahl et al., 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 quality 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 consistency. 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 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. 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

[1] Equivalently, judges convict a defendant if his unobserved 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 idea is identical.

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 more easily distinguish between high and low quality cases — increase faster than for inconsistent judges when the severity of the first-round judge (and corresponding case quality of the approved) increases. I show that the distribution of unobserved case quality can also be identified using instruments for judge leniency, which allows the construction of counterfactuals for different judge assignment mechanisms or improved judge consistency. Combining the two sources of identification, I build a structural model 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 off of patient survival measures. Also in a medical context, 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 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 making the same decision, and 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 even 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 et al., 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

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

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, but 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 relative to case quality.

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 18% of their approved choices. This 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 24.5% of IRB denials of refugee status would be overturned, rather than the 6% that is ultimately successful under the current system. This difference amounts to approximately 10,700 families over my study period.

I survey refugee lawyers about judicial quality, and validate the model by showing that survey responses are correlated with my measures of consistency and leniency. Consistency improves with judge tenure in a very interesting way. For first-round decisions, judicial consistency improves dramatically over the first year, and continues to improve (albeit at a slower rate) through at least the first ten years of experience. For second-round decisions, which typically call on a larger set of precedent, consistency improves with judge tenure at a slower rate. The biggest gains are made not after the first year, but after the tenth. I take this as evidence that longer terms for judges may improve the consistency of the entire justice system, particularly for complicated cases. I also find that consistency decreases when judge workloads are higher, particularly for judges with less than five years of experience.

I also study a reform designed to stop the government from appointing unqualified party supporters as judges. The reform, which gave a committee of legal experts veto power over candidates, reduced 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 also increased 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.

The paper proceeds in five parts. In Section 2 I present the model and discuss identification.

[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.

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 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 judges are quasi-randomly assigned in both rounds.

2.1 Model

The court receives a flow of applicants for refugee status. Each applicant i can be described by a scalar representing case strength, $r_i \sim F_r$. 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

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

where $\widetilde{\varepsilon}_{ijs} \sim G_{js}$ and $\exists x_{ijs} \in X_{ijs}$ s.t. $x_{ijs} \notin X'_{ijs}$. 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} (or equivalently, X_{ijs} shifts the distribution of r_i).

Judge *consistency* is defined by the distribution of $\varepsilon_{js}(X_{ijs})$. For perfectly consistent judges, $\varepsilon_{js}(X_{ijs}) = 0$. Then, the decision problem is non-stochastic for a given value of r_i: $P[r_i > \varepsilon_{js}(X_{ijs})] = P[U > F_r(X_{ijs}\beta + \gamma_{js})] = 1 - F_r(X_{ijs}\beta + \gamma_{js})$, 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). This definition can be expanded to compare the consistency of any two judges with the same first-round approval rate. First, note that the first-round approval rate is

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

Judge A is *comparable* to judge B if $P[r_i > \varepsilon_{A1}(X_{iAs})] = P[r_i > \varepsilon_{B1}(X_{iBs})]$. Then, he is more consistent than judge B if $\exists v$ such that:

1. $G_{A1}(v - \gamma_{A1}) = G_{B1}(v - \gamma_{B1})$
2. $\forall\ w > v, G_{A1}(w - \gamma_{A1}) \geq G_{B1}(v - \gamma_{B1})$, and for some $w > v, G_{A1}(w - \gamma_{A1}) > G_{B1}(v - \gamma_{B1})$
3. $\forall\ w < v, G_{A1}(w - \gamma_{A1}) \leq G_{B1}(v - \gamma_{B1})$, and for some $w < v, G_{A1}(w - \gamma_{A1}) < G_{B1}(v - \gamma_{B1})$

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}).

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

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

2.2 Interpretation

The innovation of this model is to separately identify judge thresholds γ_{js},the distribution of the case quality unobservables 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 $\widetilde{\varepsilon}_{ijs}$ is uncorrelated with both quality 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,

$$\widetilde{\varepsilon}_{ij1} = \bar{\varepsilon}_{ij} + \check{\varepsilon}_{ij1} \quad (4)$$

The first component, $\bar{\varepsilon}_{ij}$, 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. $\bar{\varepsilon}_{ij}$ is therefore a measure of how a judge's weighting of different facets of a case differs from the consensus.

Conversely, $\check{\varepsilon}_{ij1}$ 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 $r_i - \bar{\varepsilon}_{ij} - \check{\varepsilon}_{ij1}$, with $r_i - \bar{\varepsilon}_{ij}$ fixed and variation coming only from $\check{\varepsilon}_{ij1}$.

A natural question concerns the relative size of r_i, $\bar{\varepsilon}_{ij}$ and $\check{\varepsilon}_{ij1}$. My model identifies the relative variance of r_i versus the composite error $\bar{\varepsilon}_{ij} + \check{\varepsilon}_{ij1}$, but does not directly estimate the size of $\bar{\varepsilon}_{ij}$ versus $\check{\varepsilon}_{ij1}$. 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). 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

$$P_{jk} = P[\text{Approval by } j | \text{Approval by } k] = \frac{P[r_i > \gamma_{j2} + \bar{\varepsilon}_{ij} + \check{\varepsilon}_{ij2} \cap r_i > \gamma_{k1} + \bar{\varepsilon}_{ik} + \check{\varepsilon}_{ik1}]}{P[r_i > \gamma_{k1} + \bar{\varepsilon}_{ik} + \check{\varepsilon}_{ik1}]}$$

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 μ's and their σ's. More formally, index P_{jk} by the correlation in ideological errors $\bar{\varepsilon}_{ij}$ and $\bar{\varepsilon}_{ik}$, ρ_{jk}. It is simple to show that $P_{jk}(\rho_{jk})$ is increasing in ρ_{jk}. A reduced-form test for whether there are significant judge-pair agreements in ideology is

$$\mathbb{1}[\text{Approval by } j|\text{Approval by } k] = \beta P_{jk}(0) + \nu_{jk} + u_{ijk} \tag{5}$$

where $P_{jk}(0)$ is calculated from the model.[4] 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 $\check{\varepsilon}_{ijs}$ are larger than inter-rater differences $\bar{\varepsilon}_{ij}$. I implement this test in Section 4.11, and fail to reject the null of no judge-pair effects.

Equation 4 takes a similar form in the second stage, but in my context the decision may also be affected by new information learned. 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 an observational error $\check{\varepsilon}_{ij2}$ and an inter-rater disagreement $\bar{\varepsilon}_{ij}$ in that this new information could have the same affect on all judges. In other words, in the second round the error can be decomposed

$$\widetilde{\varepsilon}_{ij2} = \bar{\varepsilon}_{ij} + \check{\varepsilon}_{ij2} + \mathcal{I}_{i2} \tag{6}$$

The potential presence of this information complicates interpretations of the size of the distribution of $\widetilde{\varepsilon}_{ij1}$, because not all of the variation is attributable to the judge. I return to this issue in Section 3.4.

2.3 Identification

The model is identified from two different sources of variation: the random assignment of judges of varying severity, and instrumental variables. I consider each in turn.

2.3.1 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. Abstracting away from covariates X_{ij1} and substituting $\widetilde{G}_{\cdot}(r) = G_{\cdot}(r - \gamma)$ for clarity, this can be seen by noting that:

[4] 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.

$$
\begin{aligned}
&P[r > \varepsilon_{A1} \cap r > \varepsilon_{C2}] - P[r > \varepsilon_{B1} \cap r > \varepsilon_{C2}] \\
&= \int \left[\widetilde{G}_{A1}(r) - \widetilde{G}_{B1}(r)\right] \widetilde{G}_{C2}(r) f_r dr \\
&= \int_{-\infty}^{z} \left[\widetilde{G}_{A1}(r) - \widetilde{G}_{B1}(r)\right] \widetilde{G}_{C2}(r) f_r dr + \int_{z}^{\infty} \left[\widetilde{G}_{A1}(r) - \widetilde{G}_{B1}(r)\right] \widetilde{G}_{C2}(r) f_r dr \\
&\geq \int_{-\infty}^{z} \left[\widetilde{G}_{A1}(r) - \widetilde{G}_{B1}(r)\right] \widetilde{G}_{C2}(z) f_r dr + \int_{z}^{\infty} \left[\widetilde{G}_{A1}(r) - \widetilde{G}_{B1}(r)\right] \widetilde{G}_{C2}(z) f_r dr \\
&= \widetilde{G}_{C2}(z) \int \left[\widetilde{G}_{A1}(r) - \widetilde{G}_{B1}(r)\right] f_r dr \\
&= 0
\end{aligned}
\tag{7}
$$

where z is the point of single-crossing of $\widetilde{G}_{A1}$ and $\widetilde{G}_{B1}$. 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, his decisions are informative about which first-round judge has chosen higher-quality first-round claimants. The last equality comes from 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, we 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 relative to judge C.

Define judge A and B as second-round comparable if they have the same second-round approval rate conditioning 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 when going from judge D to judge C, judge A is more consistent than judge B. This can be seen by the following derivation,

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

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.

2.3.2 Instrumental variable 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, additional large-support continuous instruments are required. These instruments affect judge thresholds γ_{js} (or equivalently, shift the distribution of r) 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. A full proof of identification is in Chen et al. (2000); which I reframe in terms of my model in Appendix Section 1.1. The main identifying assumption — that the instrument affects approval rates only through thresholds γ rather than judge errors $\widetilde{\varepsilon}_{ijs}$ — is partially testable. I discuss and implement this test in Section 4.6.

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 this section by describing the IRB in enough detail to contextualize the distribution of initially denied claimants who appeal to the Federal Court. I then describe the Federal Court and the procedure the government uses to select justices for the Court.

3.1 Immigration and Refugee Board

Initial screening of inland refugee claims is conducted by 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 IRB approves about 50% of claims but between-Member variation in approval rates is large, perhaps because they are political appointees rather than professional bureaucrats. Between 2006 and 2010, the 10^{th} percentile Member approved 15.8% while the 90^{th} 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 10^{th}-90^{th}

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 implies 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. Most denied claimants file an appeal, which allows most claimants to stay in Canada until the Federal Court makes its final decision.[5]

IRB procedures for making refugee determinations were broadly consistent from 1995 until December 15, 2012, when an administrative appeal division largely 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 the identity of the potential Members before deciding whether to proceed with a single-judge panel, and 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 the bureaucrats deciding whether they would be granted 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 refugee quality 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 quality r_i to vary before and after IRPA. More details are in Section 4.4.

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 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 only because it is (imperfectly) 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.[6] If the petition for leave is approved, the claimant goes before a full judicial review (JR) hearing in front of another judge. Regardless of the first-stage outcome,

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

[6]There are two 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).

however, 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 it is relatively difficult for judges to learn about how their colleagues have ruled on similar cases.

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. Crucially, they 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 decision-maker made a "reasonable" decision (Rehaag, 2012).[7]

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 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," where they are responsible for making a determination on all cases submitted in the region during that week. There is no review of the cases before they are given to the judge, and the schedule that says which judge will be on leave duty is not public. Previous research claims that this assignment is as good as random (Rehaag, 2007); in Section 4 I show that judge characteristics are 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. 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 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 would 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

[7]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.

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 3.4.

3.3 Selection of Federal Court Justices

Federal Court justices are appointed by the Minister of Justice. For most of Canadian history the Minister has had nearly unfettered discretion over appointments and has used 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.

A major reform in 1988 reduced the discretion of the government in making appointments. The reform introduced province-level judical advisory councils (JACs) that pre-screened the list of candidates that went to the Minister for consideration. 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 federal 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.[8] The standards had some 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. The before-after comparison suffers from a lack of fully comparable data, but Russell and Ziegel (1991) report that before 1988 at least 47% of appointed judges had some involvement with the ruling Conservative party.[9] Their data comes from reports by surveyed respondents in the legal progression, and so estimates are likely biased down. For the years after the reform, Hausegger et al. (2010) cite administrative records and find that only about 30% of 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 quality of the courts may have improved as a result. I will test this hypothesis in Section 4.10.

3.4 Estimation

Fully nonparametric identification as outlined in Section 2.3 suffers from two main drawbacks. First, global identification requires special regressors with large support conditional on judge assignment. This is a very high bar. In my application, I use dummies for time of day of the hearing and day

[8] 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.

[9] 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.

of the week of the decision, which are non-continuous and do not cover the whole support. Second, the necessary deconvolutions are ill-posed inverse problems and difficult to solve in practice.

Instead, I parameterize the distributions of r_i, $\widetilde{\varepsilon}_{ij1}$ and $\widetilde{\varepsilon}_{ij2}$, generating tractable analytic expressions for approval probabilities. This allows rapid evaluation 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 larger inconsistency, ie a wider distribution of judge errors $\widetilde{\varepsilon}_{ijs}$.

As discussed in Section 3.2, the distribution of unobserved case quality 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-likelihood refugees are *not* initially granted status by the government. Referring back to Equation 1, I therefore assume that r_i is exponential-Pareto distributed with a scale parameter of 1 and a shape parameter of 1. Since there was a potential change in the distribution of case quality after changes to how the government decided refugee applications in 2002, I allow the distribution of quality to be flexibly different after the implementation of the rule changes (see Section 3.1 for more institutional details). 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. In the Appendix, I show that the distributional assumption on r_i can be relaxed by instead assuming that the distribution of r_i is a mixture of exponential-Pareto's.

In Appendix Section 1.4, I show that the probability of first-round approval when the exponential-Pareto location parameter is x_m and the scale parameter α is

$$P(r > \widetilde{\varepsilon}_{js}(X_{ijs})) = \Phi\left[\frac{\ln(x_m) - X_{ij1}\beta_1 - \gamma_{j1}}{\sigma_{j1}}\right] + e^{\alpha(\ln(x_m) - X_{ij1}\beta_1 - \gamma_{j1}) + \frac{\alpha^2\sigma_{j1}^2}{2}}\left[1 - \Phi\left(\frac{\ln(x_m) - X_{ij1}\beta_1 - \gamma_{j1}}{\sigma_{j1}} + \alpha\sigma_{j1}\right)\right] \quad (8)$$

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 for ε_{ij1} and ε_{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 1.4.

Full identification requires instrumental variables that shift judge thresholds γ but do not affect errors σ. One possible candidate would be large changes to the legal standards of review, but in this context there were no changes that had an effect on overall approval rates. Chen et al. (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 order in which decisions were made.

I instead use the timing of the decisions during the week, and of the second-round hearing during the day, as instruments. In general, I find that decisions held just before lunch and later in the week are more likely to be negative for the claimant. This matches the finding of Danziger et al. (2011), who show that Israeli judges are less likely to grant parole just before lunch. They argue that when decision-makers have made many decisions in a row, they are more likely to pick the default option. In their context, that means denying parole; in mine that means rejecting the

claimant's appeal. 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. I test the identification assumptions in Section 4.6.

In many of my specifications, I additionally allow a subset of X_{ijs}'s to affect the distribution of the error. Specifically, I model the error as distributed

$$\widetilde{\varepsilon}_{ijs}(X'_{ijs}) \sim \mathcal{N}(0, e^{\sigma_{js}+X'_{ijs}\psi}) \tag{9}$$

where $X'_{ijs} \subset X_{ijs}$. Identification using instrumental variables requires that X_{ijs} contain variables not in X'_{ijs}, so that there is identifying variation in X_{ijs} conditional on X'_{ijs}.

As discussed in Section 2.2, the size of second-round errors σ_{j2} also contains information about the informativeness of full hearings. For this reason, I focus most of my analysis on σ_{j1}. To reduce the number of parameters and improve precision, I assume that $\sigma_{j2} = \sigma_{k2}\ \forall j, k$, while allowing σ_{j1} to vary flexibly (the Appendix contains a version of the model without this restriction).

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

3.5 Data

My main data come from Federal Court case reports available on their website.[10] 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 before the implementation of the Refugee Appeal Division on June 28, 2012, a major reform that limited the jurisdiction of the FC over most appeals. 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.

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 set of IRB case files for 2006-2014 to the court records using case numbers where possible.[11] 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.[12] The average judge has 6.5 years of experience, with a maximum of 28.

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

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

[12] 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.

I exclude cases that were not perfected[13] or were unopposed by the government. I include only judges who decided cases in both the first and second round.

4 Results

4.1 Randomization tests

In Table 1 I explore whether the cases are really assigned quasi-randomly to judges. For each round, I regress claimant characteristics on judge-level mean approval rates in that round, controlling for office and year fixed effects. I also regress the characteristics on judge fixed effects and the same covariates, reporting the F-stat and p-value for the joint test of the judge fixed effects below.

The predictive power of the judges is low for the sample 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. The joint tests of the judge fixed effects are rejected slightly more than half the time, although the F-statistics are small.

Columns 1-6 are relatively straightforward outcomes: gender and country of origin. However, it is not clear how to weigh the different columns. To get around this problem, 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.

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 IRB decisions. Perhaps because of this, rates of leave-granting in the first round are low, at only 14.4%. However, there is a large amount of 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 43%. 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 factor that judges can observe, 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

[13] That is, those cases where all the paperwork was not filed on time and a decision was not made.

approval rate for both judges. The first column shows that being assigned a second-round judge who approves 10 percentage points more applicants implies a 9.3 percentage point higher chance of being approved. In the second column, *having been approved* by a 10 percentage point more lenient first-round judge gives you a 2.6 percentage point lower chance of being approved. 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 1.2, 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.[14]

Figure 1 demonstrates that there is substantial variation in judges' propensity to approve refugee claims — in the language of the model, that there is variation in γ. For evidence on variation in σ, 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 plots two pieces of reduced form evidence on the size of the variation in this ability. For each first-round judge, I take the mean approval rate of her approved claimants in the second round. Panel A displays the histogram: a 10^{th} percentile judge has 37% of his claimants ultimately approved; a 90^{th} 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 very 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. The regression coefficient is negative, but there is a large degree of dispersion in second-round approval for each first-round approval average. 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 Decision timing as instrumental variables

The main fear with using the day and times of judge decisions as instruments is that they may be affected by the underlying case quality. In particular, judges may delay difficult decisions, which would mean marginal cases would be more likely to be decided later in the week. For second-round hearings, the scheduling is done by other court officials so timing is unlikely to be correlated with case quality. In the first round, judges are given a stack of cases that they dispose of over the week, so this is a real possibility. I get around this problem by using the day of the week of the previous filing in the case as the instrument, because it both affects when the judges makes his decision (by affecting when he can look at it) and seems to be uncorrelated with active decisions by any involved parties.

[14] 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.

For first-round decisions, my data contain information on the date the decision was released. This is unfortunately not necessarily the date the decision was made. Whenever the first-round judge grants a second-round hearing, the court must hold the hearing during the 90 days after the decision is announced. Thus, the court waits to publicly release positive decisions until a judge will be available to hold a hearing. By an institutional quirk, these decisions are typically released on Fridays after judges return from a week of travels. This means that the day a leave decision was released is a low-powered and exclusion-violating instrument. I instead use the day of the week of the last filing before the decision was made as an instrument. For offices where there is no lag in scheduling hearings, having the last pre-leave filing happen on Thursday, Friday or the weekend (though weekend filings are very rare) predicts the decision will be made after Monday. As I show in Panel A Table 3, it also predicts actual approval. It is unlikely that either side is able to exploit this result through strategic filings — the last filing is usually the claimants response to the government's brief opposing leave, and must be submitted within 10 days of the government's filing. This leaves very little room for discretion, given that the identity of the judge making leave decisions is unknown to outsiders. To test this, I perform a placebo test by regressing predicted approval (where I predict approval using gender and ethnicity of the claimant) on the instrument. The coefficients are orders of magnitude smaller than the actual first stage, and statistically insignificant.

In the second round, the time and day of the week of the hearing are scheduled by court staff without detailed knowledge of the content of the cases. Thus, it is unsurprising that neither a noon hearing nor the day of the week predict *predicted* approval rates. Both are predictive, however, of actual approve. Similarly to Danziger et al. (2011), I show in Table 3 that approval rates are lower for claimants unlucky enough to have their hearing scheduled at noon. Cases heard on Wednesday or later are less likely to be approved.

4.4 Structural results

My baseline model includes in X_{ijs} controls for office of origination as well as the instrumental variables for the first-round end-of-week decision, second-round end-of-week hearing and a dummy for whether the second-round hearing was heard at lunch. Each instrument affects only the relevant round. Thresholds γ_{js} vary by judge. I allow judge error σ_{j1} to vary by by judge and not by any covariates (ie, X'_{ijs} is empty). For precision, and because some of the distribution of the second-round error is an informational shock rather than judge error, I focus most of my analysis on the first round error σ_{j1}. For precision, I therefore assume that σ_{j2} is constant.

Figure 3 plots the distribution of judge-round specific γ and σ. 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 et al. (2008). The coefficients are precisely estimated, so for most of the distribution for all the coefficients 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_1 is large relative to the distribution of refugee quality r_i, plotted in red. In Panel C, the distribution of second-round thresholds γ_2 is slightly narrower, and slightly smaller on average. The correlation between round-specific γ's is relatively high (0.46), and 80% of judges have a higher γ_{j1} than γ_{j2}. This may reflect the court's instructions to approve in the first round candidates that have an arguable case in the second round.

Panels B and D show the round-specific distributions of σ_{js}. In Panel B, the distribution of σ_{j1} is large; the tenth percentile judge judge has a σ_{j1} of 0.94 and the 90th percentile judge 2.45. This

inaccuracy makes a substantive difference in terms of who is approved by the Federal Court; pairs of judges who approve the same share of claimants disagree on at least 18% of approved claimants on average. This can be calculated by comparing approval shares for each case quality measure r_i, and assuming that for each r_i the higher-approving judge approves all the same claimants as the lower-approving judge. The true share of disagreements could be much higher. A different way of describing the size of judge-specific errors relative to the quality factor is a simple variance decomposition of the composite error $\widetilde{\varepsilon}_{ijs} - r_i$. For the baseline model, the idiosyncratic error accounts for 76% of the variance. In other words, there is an important degree of cross-judge agreement on the ranking of cases by quality, but even more judge-specific variation.

Another interesting finding from Figure 3 is that most judge's estimated σ_{j1} is smaller than the aggregate σ_2 coefficient. This is not an artifact of requiring a common σ_2 for all judges — Appendix Table A2 estimates the model without this requirement and shows a similar result. Why is this so striking? Recall from Section 2.2 that second-round residuals contain both judge error and informational shocks from what is learned at the hearing. This suggests that the distribution of judge errors $\bar{\varepsilon}_{ij} + \check{\varepsilon}_{ij2}$ — which is necessarily smaller than $\widetilde{\varepsilon}_{ij2} = \bar{\varepsilon}_{ij} + \check{\varepsilon}_{ij2} + \mathcal{I}_{i2}$ after netting out informational shocks $\mathcal{I}_{i2}$ — 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.

Although second-round judges are relatively accurate, on the aggregate first-round judges do a relatively poor job of predicting which claimants will be successful in the second round. 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 8% chance of first-round approval, but conditional on approval has a 17% chance of approval in the second round. This does not reflect selection, which is accounted for by conditioning on r. Instead, it shows that the information disclosed in the full hearing in the second round is important, and even claimants with a relatively low quality factor r_i are frequently approved. Similarly, high-r_i individuals with a 90% chance of first-round approval have only a 70% chance of approval in the second round. Integrating over the entire distribution of r, I find that on average 24.5% 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 (recall that only 14% are even given a full hearing), and underscores the importance of the additional information gained in the full hearing in terms of informing the judges decision. Although it may be surprising that the number of refugee appeals granted by the Federal Court would quadruple under a relatively small change to the system, this 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 70% of claimants in the first round, and of those 30% were approved by the second-round judge, bounding the overall approval rate in the absence of a first round at 21% (0.70×0.30). This analysis also mirrors the marginal treatment effect estimated in Appendix Section 1.2, where I show that there is relatively little variation in the MTE over the range of unobserved quality.[15]

[15]This analysis is analogous to the MTE for my model. Because the error contains transitory and idiosyncratic components, individuals are not marginal with respect to particular values of the instruments — first-round approval is always stochastic. However, my model maintains the flavor of the MTE by arranging individuals with the same quality factor by likelihood of first-round approval.

4.5 Judge randomization identification

As discussed in Section 2.3, one source of identification of σ_{j1} comes from observing whether claimants approved by a first-round judge are subsequently approved in the second round. Figure 5 shows how this intuition is reflected in the estimated model.

Judge thresholds γ_{j1} principally affect approval rates in the first round. In Panel A, 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 in the changes away from the observed value of 14%. Reassuringly, the change is monotonic and steep.

Recall from Section 2.3.1 that a main source of identification of the size of first-round judge errors 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 this moves the σ_{j1}'s away from their estimated values, this dramatically reduces the second-round approval rate.

Panel C of Figure 5 harnesses the intuition of Equation 7 more directly. The model predicts that for pairs of judges with similar first-round approval rates, the more accurate judge will have a higher second-round approval rate. 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 more model are more consistent.

In Appendix Section 1.5 I estimate a version of the model that allows σ_{j2} to also vary by judge. Although this model is less precisely estimated, I use it to demonstrate how judge randomization enables identification of second-round consistency σ_{j2}.

4.6 Instrumental variable identification

The other source of identification in the model is instrumental variables that change judge thresholds but do not judge errors $\widetilde{\varepsilon}_{ijs}$. In the first round, I use a dummy variable that predicts the decision will be made after Monday. In the second round, I use a similar end-of-week dummy for the hearing, as well as a dummy for whether the hearing was held at noon (I do not observe the day or time the decision was written after the hearing). In Section 4.3 I show that these instruments predict approval but do not predict fixed refugee characteristics associated with approval, implying relevance and exogeneity. This evidence, however, does not show that the instrument affect outcomes *only* through affecting the judge threshold $\gamma_{js} + X_{ijs}\beta_s$, rather than the size of the observational error $\sigma_{js} + X'_{ijs}\psi$. For that, I turn to an over-identification-style test in the second round, where I have two instruments. The idea is to test whether the standard deviation of the observational error varies with the instrument X'_{ijs}, using the other instrument to identify the model. Referring back to Equation 9, this amounts to testing the size and significance of ψ for specifications where X'_{ijs} includes one of the two second-round instruments. Appendix Table A3 conducts this test. It confirms that both instruments are strong, though the lunch instrument varies the decision threshold by about four times the end-of-week instrument. In terms of affecting the distribution of the error, the lunch instrument has nearly no affect, with an insignificant coefficient of -0.011 (given the log-log specification, a 1% affect). The end-of-week instrument is slightly larger, at 0.032, and marginally statistically significant (p-value 0.094). Given the strength of the lunch instrument that is identifying the coefficient ψ, this is relatively small and suggests that the second-round

threshold instruments are imperfect but close to satisfying the condition that they do not affect the observational errors. For the first round, access to only a single instrument makes this sort of test unfortunately impossible.

4.7 Observational errors and expert opinion

The judge error parameters are related to readily-observable reduced form moments in the data, and as I will show in subsequent sections change with experience, workload and judicial reform in largely predictable ways. In this section, I explore whether they are also related to lawyer's perception 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.

To measure expert opinion, I conducted a 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 1.3.

Each response was on a five-point likert scale, which I normalize by the mean and standard deviation. Table 4 describes the relationship between model coefficients and the survey results. I model the relationship as

$$\widehat{C}_{j\ell} = \beta_0 + \beta_1 \text{Favorability}_{j\ell} + \beta_2 \text{Consistency}_{j\ell} + \eta_\ell + u_{j\ell} \tag{10}$$

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.[16] To account for estimation error in the model coefficients I use Hanushek's (1974) efficient estimator.[17] 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 0.25 lower γ_1, which is about 0.08 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 0.58, or 0.28 SDs of the judge distribution. The stronger relationship is likely because second-round judge behavior is more salient for lawyers, who only appear in front of the judge in the second round.

Finally, Panel C shows the relationship between reported judge characteristics and σ_1. Reported consistency is closely related with the model estimate of σ_1; across specifications one extra SD of consistency translates to between 0.147 to 0.222 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 refugee quality and use common standards.

[16] The same regression that uses coefficients that are not adjusted for experience gives similar but less precise results.

[17] 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.

4.8 Judicial errors and experience

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 6 presents models where I allow experience to enter the judicial threshold γ_1 and the variance of the error, σ_{j1}. I parameterize experience linearly; with an indicator for more than one year of experience; and with indicators for more than 1, 5, and 10 years of experience. Column 1 shows that σ_{js} gets approximately 6% smaller with each year of experience. In Column 2, I allow experience to affect each round differentially. The effect is ten times as large in the first round, which lines up with the fact that second-round judges have published decisions to rely on for precedent.

In Columns 3-6, I discretize experience into dummy variables for more than 1 year, more than 5 years, and more than 10 years of experience. On average (in Column 5) about half of the overall improvement comes in the first year, after which the standard deviation of judicial errors drops by about 60%. There are further improvements of an additional 23 and 38% after five and ten years, respectively. In Column 6, I allow differential impacts by round. Interestingly, for first-round decisions the gains are relatively front-loaded — an 83% improvement after one year versus 40 and 89% gains over relatively longer periods of 5 and 10 years. For second-round decisions, however, improvement is slower (the overall improvement is about 10 times smaller) and back-loaded — the largest gains come after 10 years. Recall that second-round decisions are somewhat more complicated — the justice conducts a full hearing with lawyers from both parties before making her decision, as opposed to reviewing documents from the IRB's original determination in the first round. The existing precedent in the second round is more directly applicable, and absorbing it all may take time. Thus, one likely reason that improvement is slower and more back-loaded for second-round decisions is that there is simply more to learn. One obvious implication of this is that it may be efficient to assign more experienced judges to second-round decisions. In Appendix Table A7, I estimate a model that allows for linear year controls in judge consistency. The results are somewhat less precise but substantively similar.

4.9 Judicial accuracy and workload

There is significant over-time variation in the workload judges face. Does this affect decision consistency? I calculate monthly log workload as the number of leave cases a judge is assigned in a given month, and in Table 7 test whether it affects errors. This is an imperfect measure of workload, since judges are also responsible for non-refugee cases. However, because this model accounts for judge-specific consistency in σ_{js}, these estimates do *not* reflect time-invariant selection of more- or less-consistent judges into refugee work. However, to the extent that judges with higher refugee caseloads likely have lower non-refugee caseloads, these estimates are likely biased towards zero.

In column 1, I allow workload to affect both first- and second-round errors. The effect is large; 10% more cases increases the standard deviation of judge error σ_j by 0.9%. Restricting caseload to only affect first-round errors, the effect increases to 1.6% (column 2). This is consistent with first-round decisions being made more instinctively, and thus more sensitive to environmental factors. In column 3, I add controls for experience, which slightly reduces the size of the coefficient. In Column 4, I distinguish between experienced and inexperienced judges. As one might expect, the effect becomes smaller with experience. For judges with less than 5 years of experience, a 10%

increases in the number of cases reduces consistency σ_{j1} by 1.2%. Judges with more than five years of experience, however, have their consistency reduced by only 0.5%. Taken together, these results suggest that excess cases should be disproportionately assigned to experienced judges to a greater extent than they currently are.

4.10 Judge selection reform and judicial quality

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 new policy reduced the number of judges with ties to the ruling party. In this section, I provide evidence that the reform was also successful in improving the consistency σ_{js} of the judges.

Table 5 presents a regression of σ_{j1} on a dummy for whether the judge was appointed before the reform. I weight the regressions to account for the estimation error in the dependent variable (Hanushek, 1974), and in my preferred specifications control for party of appointment. Because the reform took place seven years before the start of my sample, the pre-reform judges are mechanically more experienced. Since more experienced judges are more consistent (have lower σ_{js}), this works against finding that the reform improved judge consistency.[18]

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 log 0.54 after the reform. 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 much larger once the model properly accounts for differing experience among pre- and post-reform judges, with log σ_{j1} dropping by 1.1 (relative to a mean of 0.75) for judges appointed after the reform.

Both of these affects are large. 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 those candidates would otherwise have been appointed. 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).

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.

[18] 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} 0.22 lower (standard error 0.15) than non-promoted judges.

4.11 Ideology versus observational errors

In my model, judicial inconsistency is composed of 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.

In my data, occasionally the same judge is assigned to both the first- and second-round decision. As I discuss in Section 2.2, 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. This correlation is estimated to be positive and quite large, at 0.33 (SE=0.01). In a reduced-form sense, this is because second-round justices disproportionately approve claimants that they approved in the first round, above and beyond what would be expected by their overall first- and second-round approval rates. This could result 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 between $\widetilde{\varepsilon}_{ij1}$ and $\widetilde{\varepsilon}_{ij2}$ implies that at least one of these factors is large. In Section 2.2 I describe a test of the relative size of these errors. If the difference is ideological inter-rater inconsistency, then it is likely that other judges share the same weighting of different aspects of cases. Then, judge-pair fixed effects would predict second-round approval beyond the model-based probabilities.

I implement this test in Table 8. The left three 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 three columns ignore ordering.

In Column 1, without conditioning on the model probabilities, the judge-pair fixed effects are jointly significant. However, conditioning on the the model probability in Column 2 and additionally on the model covariates in Column 3, the F-stat for the joint test of the judge-pair fixed effects drops to 1.05. This corresponds to an asymptotic p-value of 0.122. Since many of the judge-pair cells are small when not ignoring ordering, I follow Abrams et al. (2012) and bootstrap the distribution of the null. Compared to the bootstrap distribution, an F-stat of 1.05 gives a p-value of 0.464.

As a descriptive analysis, I calculate the Empirical Bayes judge-pair means.[19] This confirms the F-stat result: the standard deviation of the judge-pair means is only 0.013 relative to a mean approval rate of 0.44.

The rightmost three columns, where I pool judge pairs by disregarding ordering, tells a similar story: once I have conditioned on the model probabilites, knowing the actual identity of the judges does not increase predictive power. The judge-pair cells are larger, so in this specification there is very little difference between the asymptotic and bootstrapped p-values.

The results in this section are a joint test of both the strength of the ideology correlations ρ_{jk} 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 $\bar{\varepsilon}_{ik}$ is high but all the ρ_{jk}s are low, because that would require that $\bar{\varepsilon}_{ik}$ is a 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, there are simply not enough aspects of the cases that each judge could consistently weight a different one of them highly than all the other judges.

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

I therefore take this as evidence that the judge errors $\widetilde{\varepsilon}_{ijs}$ are mostly composed of idiosyncratic observational errors, rather than extra dimensions of unobserved refugee quality.

4.12 Optimal judge allocation

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

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 lawyers for both sides, time to prepare for the hearing and time to write the decision — about ten times as long as a first-round decision. Heuristically, 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. However, it is not obvious from the reduced-form data what that 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 posterior case quality 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. To allow this, I use the model with individually-varying σ_{js}'s detailed in Appendix Section 1.5. Interestingly, there is a slight negative correlation between estimated σ_{j1} and σ_{j2}, indicating that judges are already somewhat specialized towards first- or second-round cases. Second, first-round judges should be made more strict to improve the posterior case quality of claimants approved in the third round. Third, second-round judges should approve a higher share of cases so that the overall approval remains the same with lower first-round approval rates.

In Figure 6, I conduct exactly this maximization. I find that overall workload would be reduced by 18% (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 all three kinds of savings. 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), but their σ_{j1} are lower, indicating more consistent decisions (Panel B). In the second round, judges are much more lenient, as evinced by the lower thresholds γ_{j2} (Panel C). Panel D shows that they are also more consistent, meaning they are more likely to approve high-r_i claimants. This keeps the posterior distribution of quality higher than in baseline.

5 Conclusion

Much research has focused on non-relevant factors that affect changed judge behavior: the decision in the previous case (Chen et al., 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). In this paper I develop a simple model where approve all candidates with a case strength larger than a judge-specific

threshold. Judges observe case quality with some error, which generates inconsistencies across judges in which claimants they approve, even between judges who approve the same share of cases. I show that this model is identified in two-stage judicial processes by a combination of across-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 instrumental variables that shift judge thresholds without affecting errors.

I implement the model using data from judicial review of Canadian refugee claims, and validate its measure of consistency against a survey of refugee lawyers. Although the judges of the Federal Court are experts in refugee cases, there are relatively high levels of inconsistency. For first-round judges who approve the same share of cases, I bound the share of approved claimants they disagree on to at least 18%. However, judicial consistency improves dramatically with experience, particularly over the first year. Judges — and particularly inexperienced ones — are more consistent when they have a smaller workload. Reforms in the late 1980s design to stop the government from appointing unqualified party supporters reduced judicial errors, 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 the initially-denied refugee claimants who appeal the decision. Future work should determine whether the results hold in criminal courts and other decision-making institutions such as the Social Security Administration. 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. I find that first-round judges reject many claimants who might be successful in the second stage — if first-stage approval became automatic, 24.5% of all claimants would have their cases returned to the government for redetermination. Over the 17 years from 1995 that comprise my study period, that difference amounts to approximately 10,700 families.

References

Abaluck, J., Agha, L., Kabrhel, C., Raja, A., Venkatesh, A., et al. (2016). The determinants of productivity in medical testing: Intensity and allocation of care. *American Economic Review*, 106(12):3730–3764.

Abrams, D. S., Bertrand, M., and Mullainathan, S. (2012). Do judges vary in their treatment of race? *The Journal of Legal Studies*, 41(2):347–383.

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

Alesina, A. F. and Ferrara, E. L. (2011). A test of racial bias in capital sentencing. Technical report, National Bureau of Economic Research.

Anwar, S. and Fang, H. (2006). An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *The American economic review*, 96(1):127–151.

Bhuller, M., Dahl, G. B., Løken, K. V., and Mogstad, M. (2016). Incarceration, recidivism and employment. Technical report, National Bureau of Economic Research.

Bureau of Justice Statistics (2006). Examining the work of state courts.

Canes-Wrone, B., Clark, T. S., and Kelly, J. P. (2014). Judicial selection and death penalty decisions. *American Political Science Review*, 108(01):23–39.

Card, D., Mas, A., Moretti, E., and Saez, E. (2012). Inequality at work: The effect of peer salaries on job satisfaction. *The American Economic Review*, 102(6):2981–3003.

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

Chen, D. L., Moskowitz, T. J., and Shue, K. (2016). Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *The Quarterly Journal of Economics*, 131(3):1181–1242.

Chen, X., Heckman, J. J., Vytlacil, E., 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*, number 1567. Econometric Society.

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

Dahl, G. B., Kostol, A. R., and Mogstad, M. (2013). Family welfare cultures. Technical report, National Bureau of Economic Research.

Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892.

Dauvergne, C. (2003). Evaluating canada's new immigration and refugee protection act in its global context. *Alta. L. Rev.*, 41:725.

Delaigle, A., Hall, P., and Meister, A. (2008). On deconvolution with repeated measurements. *The Annals of Statistics*, pages 665–685.

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

Eren, O. and Mocan, N. (2016). Emotional judges and unlucky juveniles. Technical report, 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*.

Fischman, J. B. (2013). Measuring inconsistency, indeterminacy, and error in adjudication. *American Law and Economics Review*, 16(1):40–85.

Frakes, M. D. and Wasserman, M. F. (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*, (0).

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

Glaze, L. E. and Parks, E. (2011). Correctional populations in the united states, 2011. *Population*, 6(7):8.

Grant, A. G. and Rehaag, S. (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(2):66–67.

Hausegger, L., Riddell, T., Hennigar, M., and Richez, E. (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(3):633–659.

Heckman, J. J. and Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3):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(2):99–116.

Loeffler, C. E. (2013). Does imprisonment alter the life course? evidence on crime and employment from a natural experiment. *Criminology*, 51(1):137–166.

Maestas, N., Mullen, K. J., and Strand, A. (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(3):205–228.

McKelvey, S. (1985). The appointment of judges in canada. Technical report, Canadian Bar Association.

Mueller-Smith, M. (2014). The criminal and labor market impacts of incarceration. *Unpublished Working Paper*.

Partridge, A. and Eldridge, W. B. (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. Technical report, Pew Research Center.

Porta, R. L., Lopez-de Silanes, F., Shleifer, A., and Vishny, R. W. (1998). Law and finance. *Journal of political economy*, 106(6):1113–1155.

Rehaag, S. (2007). Troubling patterns in canadian refugee adjudication. *Ottawa L. Rev.*, 39:335.

Rehaag, S. (2012). Judicial review of refugee determinations: The luck of the draw?

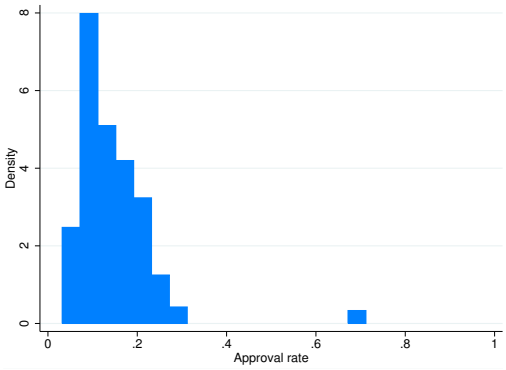
Russell, P. H. and Ziegel, J. S. (1991). Federal judicial appointments: An appraisal of the first mulroney government's appointments and the new judicial advisory committees. *The University of Toronto Law Journal*, 41(1):4–37.

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

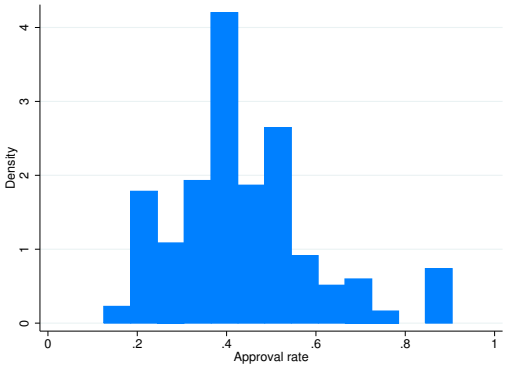
6 Figures

Figure 1: Approval rates by judge

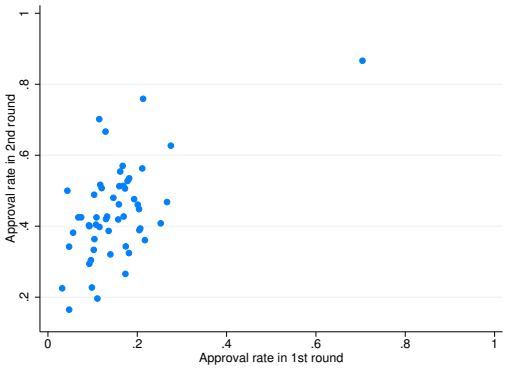
(a) First round



(b) Second round



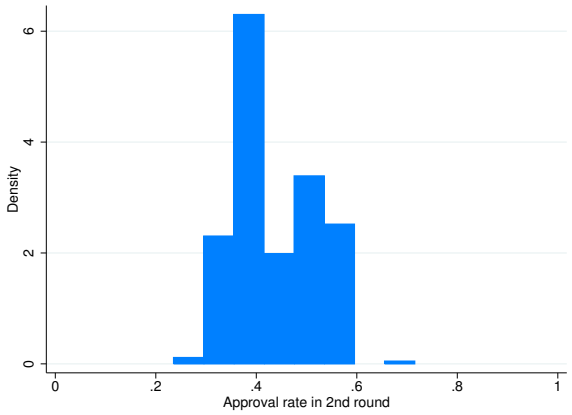
(c) Second vs. first round approval rates



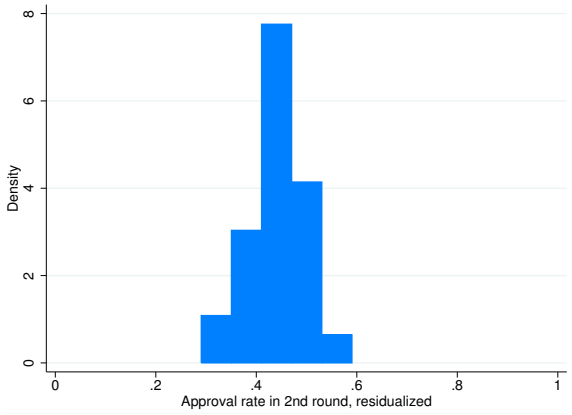
Panel A and B contain histograms of approval rates by judge for the first and second round, respectively. Both are weighted by the number of observations per judge. Panel C contains the scatter plot of judge-level first- and second-round approval rates. The correlation is 0.57, and 0.40 without the outlier.

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

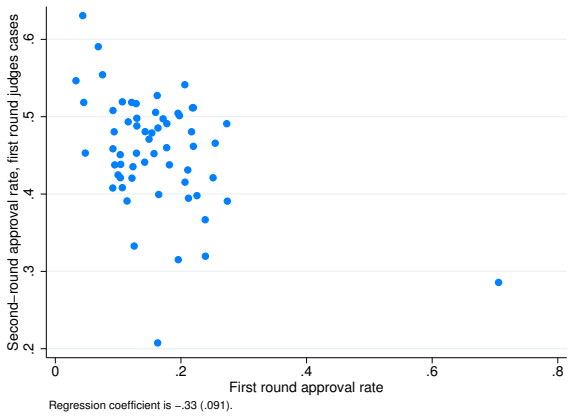
(a) Second-round approval



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



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

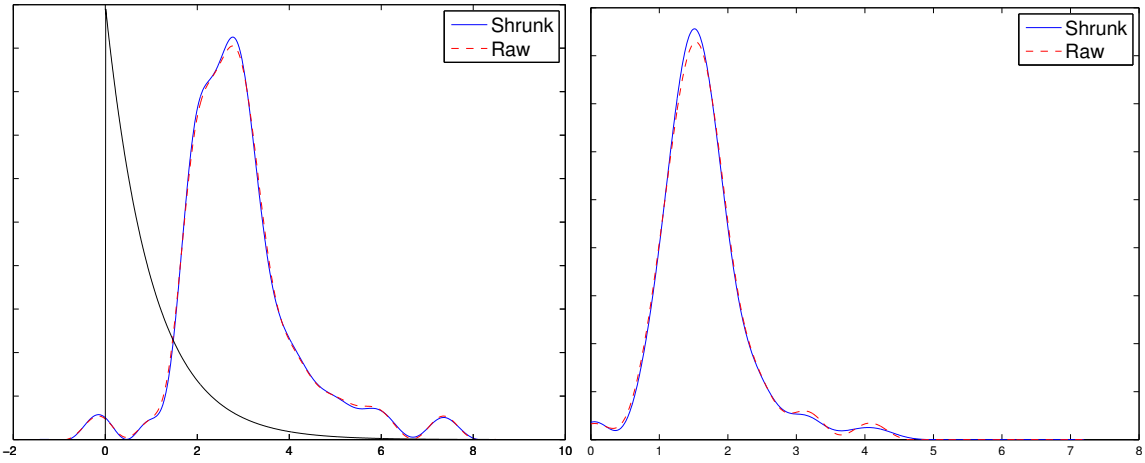


Panel A contains a histogram of second-round approval rates for the cases approved by the first-round judge. Higher approval rates suggest that the first-round judge did a better job of selecting claimants with a high probability of success in the second round. Panel B contains the same histogram, after residualizing out first- and second-round judge approval rates. Panel C plots second round approval rates for the claimants approved by each first round judge, plotted against the judge's first-round approval rates. For both, second-round judge approval rates are residualized out.

Figure 3: Distribution of judge coefficients

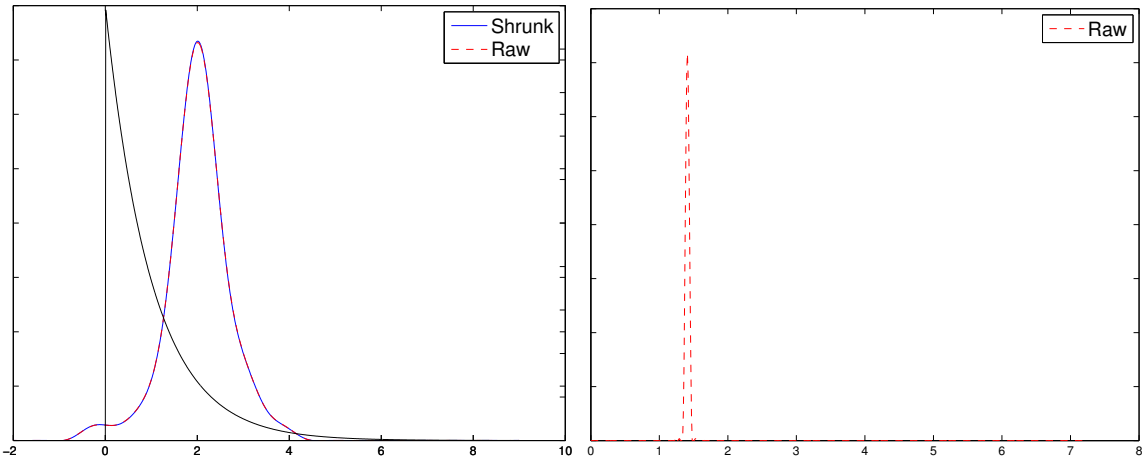
(a) Threshold γ_1, first round

(b) Observational error σ_1, first round



(c) Threshold γ_2, second round

(d) Observational error σ_2, second round



Each panel contains the density of the raw and shrunken estimates of the judge thresholds γ_1 and γ_2, and the errors 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 (Delaigle and Meister, 2008).

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

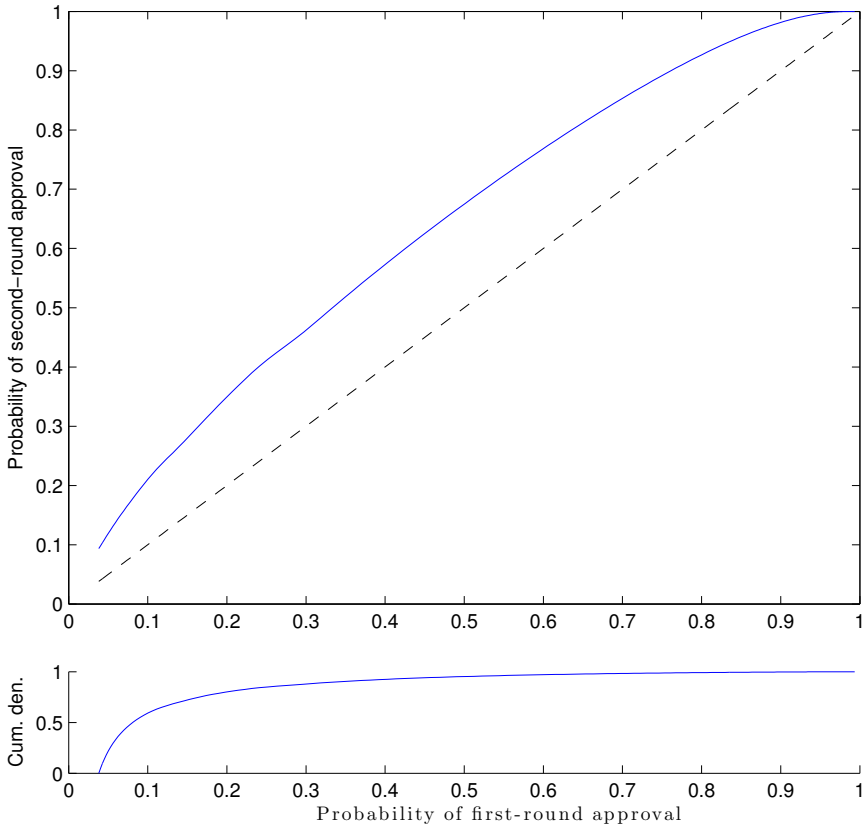
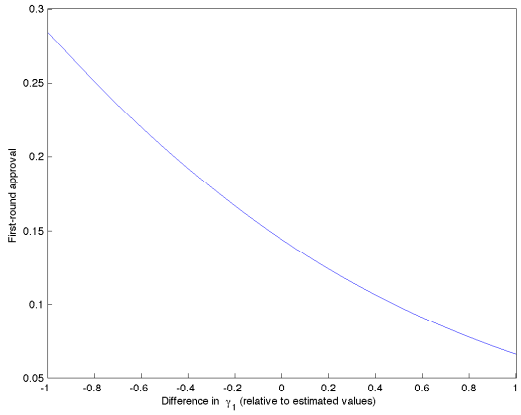


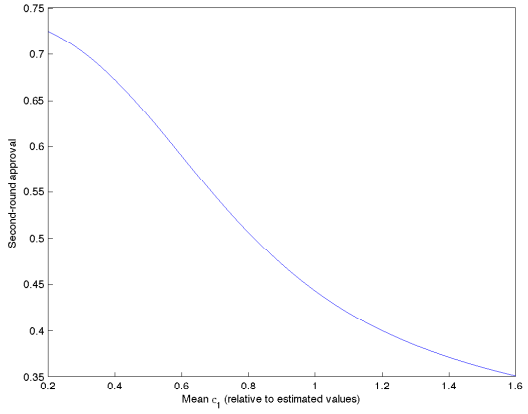
Figure plots first-round approval probability against second-round approval probability conditional on first-round approval for each value of case strength r_i. Secondary graph displays cumulative density of first-round approval. Black dotted comparison line marks out 45°.

Figure 5: Identification intuition

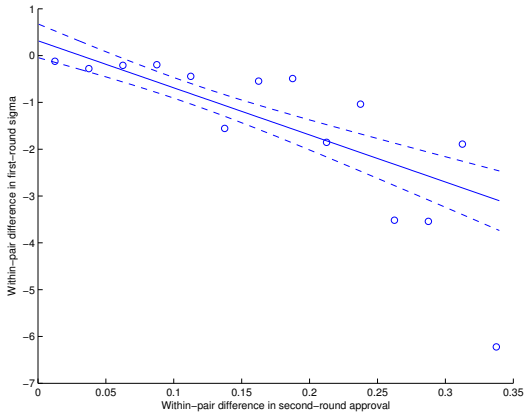
(a) First-round approval as function of first-round threshold γ_1



(b) Second-round approval as function of mean first-round error σ_1



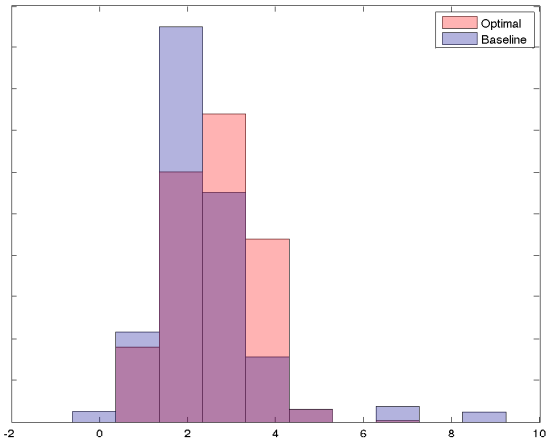
(c) Difference in first-round error σ_1 as function of difference in second-round approval for judge pairs



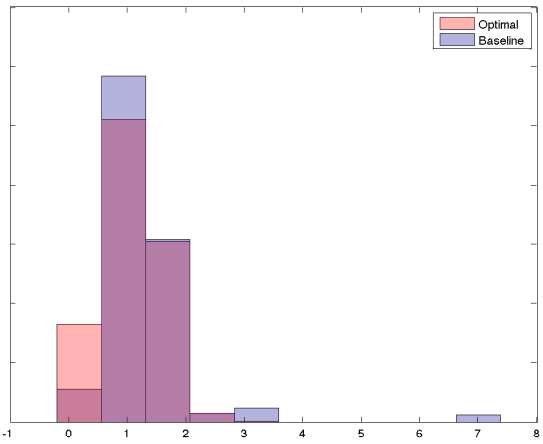
Panel A contains model estimates of the first-round approval probability as a function of deviation of threshold γ_1. Panel B contains model estimates of second-round approval as a function of mean first-round error relative to the estimated value — higher errors make second-round approval less likely. 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 $\widehat{\sigma}_1$.

Figure 6: Optimal allocation of judges

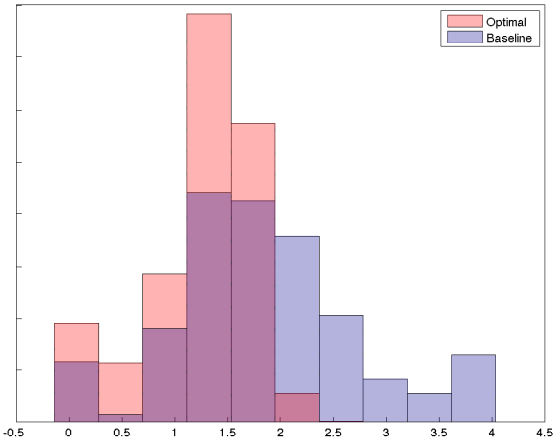
(a) Threshold γ_1, first round



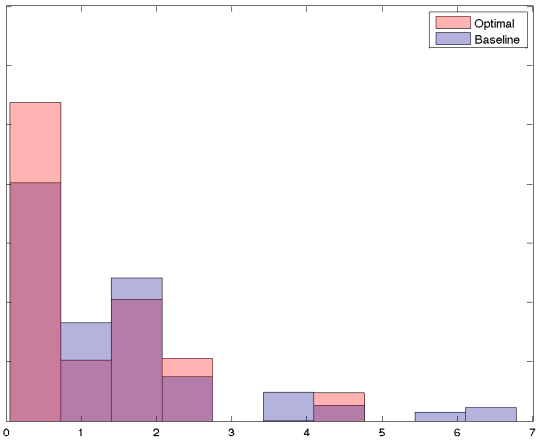
(b) Observational error σ_1, first round



(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 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 18%.

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							
Judge mean approval	-0.037	-0.018	-0.008	0.047	0.008	0.000	
	(0.024)	(0.025)	(0.036)	(0.029)	(0.015)	(0.002)	
F-stat (judge FEs)	0.88	1.69	1.40	1.30	1.68	1.54	
Prob	0.71	0.00	0.03	0.08	0.00	0.01	
Observations	20166	20166	20166	20166	19805	20166	
Panel B: Second round judges							
Judge mean approval	-0.008	0.020	-0.059	0.064	0.011	0.002	0.009
	(0.040)	(0.033)	(0.054)	(0.058)	(0.016)	(0.002)	(0.021)
F-stat (judge FEs)	1.31	0.83	1.24	1.65	0.83	1.20	2.39
Prob	0.08	0.78	0.13	0.00	0.79	0.17	0.00
Observations	4062	4062	4062	4062	3905	4062	4030

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 and continent of origin. F-stats come from separate regression of outcome on judge fixed effects. Standard errors clustered at the judge level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Second-round approval on mean judge approval

	(1)	(2)	(3)
Mean second round approval, exclusive	0.943***		0.958***
	(0.0267)		(0.0239)
Mean first round approval, exclusive		-0.258***	-0.312***
		(0.0512)	(0.0423)
Observations	8446	8480	8446

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

Table 3: Placebo tests and first stage for instrumental variables, 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				
Noon hearing	-0.001	-0.001	-0.078***	-0.075***
	(0.001)	(0.001)	(0.022)	(0.023)
Wed-Fri hearing	0.001	0.001	-0.022*	-0.022*
	(0.001)	(0.001)	(0.012)	(0.012)
Controls	No	Yes	No	Yes
Observations	8446	8446	8446	8446

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: Model coefficients on survey responses

	(1)	(2)	(3)
Panel A: γ_1 (mean=2.96, SD=1.32)			
Favorability, SD	-0.268*		-0.253
	(0.149)		(0.159)
Consistency, SD		-0.140**	-0.040
		(0.068)	(0.070)
Observations	182	182	182
Panel B: γ_2 (mean=2.09, SD=1.28)			
Favorability, SD	-0.524***		-0.580***
	(0.177)		(0.177)
Consistency, SD		-0.065	0.156**
		(0.101)	(0.075)
Observations	182	182	182
Panel C: σ_1 (mean=2.1, SD=1.75)			
Favorability, SD	0.144		0.220*
	(0.113)		(0.122)
Consistency, SD		-0.147*	-0.222**
		(0.082)	(0.101)
Observations	182	182	182

Estimated with Hanushek (1974) correction for estimated dependent variable. All models include respondent fixed effects. Standard errors clustered at the judge level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Judge inconsistency σ_j by appointment before and after reform

	Baseline			Experience ctrl in σ		
	(1)	(2)	(3)	(4)	(5)	(6)
After reform (=1)	-0.389**	-0.534*	-0.537*	-1.257***	-1.198***	-1.098***
	(0.193)	(0.291)	(0.283)	(0.298)	(0.381)	(0.383)
Liberal appointee (=1)			0.0997			-0.181
			(0.135)			(0.175)
Year appointed	No	Yes	Yes	No	Yes	Yes
Dependent mean	1.53	1.53	1.53	0.75	0.75	0.75
N judges	53	53	53	53	53	53

Estimated with Hanushek (1974) correction for estimated dependent variable. Left three columns use baseline model estimates that do not account for judge experience. Right three columns allow consistency σ_{j1} to vary with experience (by dummy variables for more than one, more than five, and more than 10 years), and adjust all judges to have six years of experience for comparability. Robust standard errors in parentheses and clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Judge inconsistency σ_j and experience

	(1)	(2)	(3)	(4)	(5)	(6)
Coefficients ψ in log judge observational error, σ_1						
Experience	-0.063^{***}					
	(0.002)					
Experience (1st round)		-0.150^{***}				
		(0.004)				
Experience (2nd round)		-0.016^{***}				
		(0.002)				
Experience $>$ 1 year			-0.830^{***}		-0.581^{***}	
			(0.023)		(0.122)	
Experience $>$ 5 years					-0.228^{***}	
					(0.032)	
Experience $>$ 10 years					-0.382^{***}	
					(0.026)	
Experience $>$ 1 year (1st round)				-1.214^{***}		-0.831^{***}
				(0.011)		(0.017)
Experience $>$ 5 years (1st round)						-0.405^{***}
						(0.018)
Experience $>$ 10 years (1st round)						-0.885^{***}
						(0.011)
Experience $>$ 1 year (2nd round)				-0.152^{***}		-0.124^{***}
				(0.028)		(0.028)
Experience $>$ 5 years (2nd round)						0.018
						(0.024)
Experience $>$ 10 years (2nd round)						-0.243^{***}
						(0.013)
SD of γ_1	0.884	0.916	1.120	0.952	0.956	1.069
SD of σ_1	1.164	2.098	1.604	1.753	1.596	1.926

Reports coefficents for decision model $\mathbb{1}[r_i > X_{ijs}\beta_s + \gamma_{js} + \widetilde{\varepsilon}_{ijs}(X'_{ijs})]$, $\widetilde{\varepsilon}_{ijs}(X'_{ijs}) \sim \mathcal{N}(0, e^{\sigma_{js} + X'_{ijs}\psi})$. All models include controls for office and time/date of decision in β, and allow the parameters of the Pareto distribution of r_i to vary flexibly 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 7: Judge inconsistency σ_j and workload

	(1)	(2)	(3)	(4)
Coefficients ψ in log judge observational error, σ_1				
Log workload (both rounds)	0.092***			
	(0.011)			
Log workload (first round)		0.164***	0.117***	
		(0.011)	(0.010)	
Log workload (first round, $\leq$ 5 years experience)				0.121**
				(0.050)
Log workload (first round, $>$ 5 years experience)				0.052**
				(0.025)
Experience control	No	No	Yes	Yes

Reports coefficients for decision model $\mathbb{1}[r_i > X_{ijs}\beta_s + \gamma_{js} + \widetilde{\varepsilon}_{ijs}(X'_{ijs})]$, $\widetilde{\varepsilon}_{ijs}(X'_{ijs}) \sim \mathcal{N}(0, e^{\sigma_{js} + X'_{ijs}\psi})$. Worload measured as log of number of first-round cases. All models include controls for office and time/date of decision in β, and allow the parameters of the Pareto distribution of r_i to vary flexibly 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 8: Second-round outcome on model approval probability and judge-pair FEs

	Judge-pair round FEs			Judge-pair FEs		
	(1)	(2)	(3)	(4)	(5)	(6)
Model approval probability		1.037***	0.885***		0.985***	0.978***
		(0.201)	(0.270)		(0.0493)	(0.0511)
Model controls	No	No	Yes	No	No	Yes
Mean approval	0.44	0.44	0.44	0.44	0.44	0.44
F-stat for judge pairs	1.70	1.04	1.05	1.58	1.00	1.01
Prob	0.000	0.163	0.122	0.000	0.472	0.417
BS prob	0.000	0.472	0.464	0.000	0.411	0.512
SD of judge-pair EB means	0.137	0.013	0.014	0.101	0.002	0.002
Observations	8196	8196	8196	8196	8196	8196

Regresses second-round approval on model-predicted likelihood of approval and judge-pair fixed effects. Left three columns construct judge-pair FEs accounting for order of assignment; right three columns ignore this distinction. Model controls include office of origination, day of week of first-round pre-leave filing, day of week of second-round hearing, and a dummy variable for a noon second-round hearing. Standard errors clustered at the judge level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

1 Online Appendix for *Judicial Errors: Evidence from Refugee Appeals*

1.1 Details on instrumental variable 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., 200), where the "duration" is the amount of time until a judge rejects the applicant's case (this "time" is capped at 2, of course). Nonparametric identification requires:

1. r_i and $\widetilde{\varepsilon}_{ij}$ are independent and have a median of zero with finite unknown variances.
2. $X_{ij1}\beta_1|\gamma_j$ is continuous with large support.
3. $X_{iks}\beta_2|X_{iks}\beta_1, \gamma_j, \gamma_k$ 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. Rewriting Equation 1, in each stage an individual is approved if

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

where H_{js} is the distribution of $\eta_{ks} = \widetilde{\varepsilon}_{ijs} - r_i$, the composite error of the refugee-level equality variable r_i and the case-judge idiosyncratic error $\widetilde{\varepsilon}_{ijs}$. 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 enters the distribution H_{js}, and thus the judge effect γ_j cannot be used for identification. Instead, X_{ij1} traces out the distribution of H_{js}, which is why Assumption 2 calls for large support conditional on judge assignment.

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

$$\lim_{X_{ij1}\beta_1 \to -\infty} \mathbb{1}[-X_{ij2}\beta_2 - \gamma_{j2} > \widetilde{\varepsilon}_{ij2} - r_i | - X_{ij1}\beta_1 - \gamma_{j1} > \widetilde{\varepsilon}_{ij1} - r_i] = H_{j2}(-X_{ij2}\beta_2 - \gamma_{j2}) \tag{2}$$

so β_2 and H_{k2} are similarly identified to scale. By Assumption 4, this scale is the same. Then, similarly to Chen et al. (2000), the variances of r_i, $\widetilde{\varepsilon}_{ij1}$ and $\widetilde{\varepsilon}_{ij2}$ are identified from the variance of the first and second round residuals and their covariance. Finally, deconvolution recovers the distributions G_{j1}, G_{j2}, and F_r.

1.2 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. The simplest 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

$$\Delta^{MTE}(u_D) = E[Y_1 - Y_0 | U_D = u_d] \tag{3}$$

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

$$D* = P(Z) \geq U_D \tag{4}$$

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 initial 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 0.71. 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 3$^{\text{rd}}$ to the 28$^{\text{th}}$ 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 40% chance of approval in the second round.

1.3 Survey questions

As I discuss in Section 4.7, 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 some 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 the 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?

   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?

[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 polling response rates, which are below 10% (Pew, 2012). However, it is significantly lower than 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), 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 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 2.2. 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 4, where I include only the first two questions. I discuss these in Section 4.7. 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 refugee 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.

1.4 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 calculation of first-round approval probabilities, then the second-round probabilities.

1.4.1 First round approval

$$\begin{aligned} P(r - \varepsilon_1 > 0) &= \int_0^{\infty} P(r > \tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_1) \\ &= \int_0^{x_m} P(r > \tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_1) + \int_{x_m}^{\infty} P(r > \tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_1) \end{aligned} \tag{5}$$

The first term in Equation 5 can be shown to be equal to $\Phi[\frac{\ln(x_m)-\mu_1}{\sigma_1}]$. Then,

$$\begin{aligned} \int_{x_m}^{\infty} P(r > \tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_1) &= \int_{x_m}^{\infty} \frac{x_m^{\alpha}}{\tilde{\varepsilon}_1^{\alpha}} \phi\left(\frac{\ln(\tilde{\varepsilon}_1) - \mu_1}{\sigma_1}\right) \frac{1}{\sigma_1 \tilde{\varepsilon}_1} d\tilde{\varepsilon}_1 \\ &= x_m^{\alpha} \int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}}^{\infty} e^{-\alpha(\sigma_1 y + \mu_1)} \phi(y) dy \\ &= x_m^{\alpha} e^{-\alpha\mu_1} \int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}}^{\infty} e^{-\alpha\sigma_1 y - \frac{y^2}{2}} dy \\ &= x_m^{\alpha} e^{-\alpha\mu_1 + \frac{\alpha^2\sigma_1^2}{2}} \int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}}^{\infty} e^{-\frac{1}{2}(y+\alpha\sigma_1)^2} dy \\ &= x_m^{\alpha} e^{-\alpha\mu_1 + \frac{\alpha^2\sigma_1^2}{2}} \int_{\frac{\ln(x_m)-\mu_1}{\sigma_1} + \alpha\sigma_1}^{\infty} e^{-\frac{1}{2}y^2} dy \\ &= x_m^{\alpha} e^{-\alpha\mu_1 + \frac{\alpha^2\sigma_1^2}{2}} \left[1 - \Phi\left(\frac{\ln(x_m) - \mu_1}{\sigma_1} + \alpha\sigma_1\right)\right] \end{aligned} \tag{6}$$

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

1.4.2 Approval in both rounds

$$\begin{aligned} P(r > \varepsilon_2 \cap r > \varepsilon_1) &= \int_0^\infty \int_0^\infty P(r > \tilde{\varepsilon}_2 | r > \tilde{\varepsilon}_1) P(r > \tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_2) \\ &= \int_0^\infty \int_0^\infty P(r > \tilde{\varepsilon}_2 | r > \tilde{\varepsilon}_1) P(r > \tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_2) \end{aligned} \tag{7}$$

The terms inside the integrals can be rewritten

$$P(r > \tilde{\varepsilon}_1) = \mathbb{1}[\tilde{\varepsilon}_1 < x_m] + \mathbb{1}[\tilde{\varepsilon}_1 \geq x_m] \frac{x_m^\alpha}{\tilde{\varepsilon}_1^\alpha} \tag{8}$$

and

$$\begin{aligned} P(r > \tilde{\varepsilon}_2 | r > \tilde{\varepsilon}_1) = &\mathbb{1}[\tilde{\varepsilon}_1 < x_m] \left[\mathbb{1}[\tilde{\varepsilon}_2 < x_m] + \mathbb{1}[\tilde{\varepsilon}_2 \geq x_m] \frac{x_m^\alpha}{\tilde{\varepsilon}_2^\alpha} \right] + \\ &\mathbb{1}[\tilde{\varepsilon}_1 \geq x_m] \left[\mathbb{1}[\tilde{\varepsilon}_2 < \tilde{\varepsilon}_1] + \mathbb{1}[\tilde{\varepsilon}_2 \geq \tilde{\varepsilon}_1] \frac{\tilde{\varepsilon}_1^\alpha}{\tilde{\varepsilon}_2^\alpha} \right] \end{aligned} \tag{9}$$

Subbing in to Equation 7 and expanding the integrals,

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

Further separate the integrals into four components:

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

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

$$\int_{x_m}^\infty \int_0^{\tilde{\varepsilon}_1} \frac{x_m^\alpha}{\tilde{\varepsilon}_1^\alpha} dF(\tilde{\varepsilon}_2) dF(\tilde{\varepsilon}_1) \tag{12}$$

$$\int_{x_m}^\infty \int_{\tilde{\varepsilon}_1}^\infty \frac{x_m^\alpha}{\tilde{\varepsilon}_2^\alpha} dF(\tilde{\varepsilon}_2) dF(\tilde{\varepsilon}_1) \tag{13}$$

These four equations (10-13) are all simple because the distribution of a Pareto-distributed random variable conditional on being larger than a given threshold is itself Pareto. 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.

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

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

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

1.5 Identification intuition for full model

My main analyses use a version of the model that fixes σ_{j2} to be the same for all judges. This is partially to increase precision, but also because the second-round error $\widetilde{\varepsilon}_{ij2}$ reflects both judge error and informational shocks from the second-round hearing. Variation in it does not reflect only variation across judges, but also the informational environment.

However, for the exercise of optimally allocating judge pairs to cases in Section 4.12, I estimate a model with no cross-judge restrictions on σ_{j2}. This allows more efficient substitution of judges towards the rounds where they are more consistent, and more accurately reflects the gains from optimal allocation. In this section I present the estimates for the unrestricted model, and explain how the judge randomization identification works in this context.

Figure A2 displays the coefficients. Unlike in the baseline model, the distribution of raw coefficients differs from the shrunken distribution that accounts for measurement error (via the deconvolution procedure of Delaigle and Meister (2008)). This is because all coefficients are estimated with substantially more measurement error. Nonetheless, the distribution of the shrunken coefficients looks similar to the baseline model. The main difference is in Panel D, which contains estimates for second-round consistency σ_{j2}. Instead of a point mass at 1.4 as in the baseline model, the unrestricted model has a spike in mass at about 0.75, followed by a long tail of second-round inconsistent judges.

Figure A3 contains the identification explanation analogous to that in the main paper. Similarly to Figure 5, increasing the first-round threshold reduces first-round approval rates, and increasing aggregate first-round inconsistency σ_{j1} reduces second-round approval. Matching pairs of first-round judges with similar first-round approval rates, judges who's approved claimants were more likely to be approved in the second round have higher estimated first-round consistency (lower σ_{j1}). Unlike in the baseline model, this allows σ_{j2} to vary across judges, which allows a demonstration of how the coefficients match the judge-randomization logic of Section 2.3.1. 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. I then show that second-round approval rates conditional on first-round approval by a less lenient judge will be higher for the more consistent judge. Panel D shows how the estimated model reflects this logic. Fortunately for identification, 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 based on approval rates conditional on first-round approval by the outlier judge, requiring that the judge has more than 10 cases and the approval rate be within 5%. In Panel D I display a scatter plot of the difference in estimated second-round inconsistency σ_{j2} and the difference in second-round approval rates conditional on first-round approval by another judge. As expected, the larger the difference in approval rates, the larger the difference in estimated σ_{j2}, with higher approval rates for the comparison judge corresponding to a lower σ_{j2}.

1.6 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. 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 a judge 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. Mueller-Smith shows how this problem can be circumvented with a LASSO first stage to select the instruments (in his case, judge and prosecutor effects interacts with defendant characteristics and crime type) with the best predictive power.

Mueller-Smith's solution depends on observing covariates that explain heterogeneous judge behavior. In this section, I use results from Klein (2010) to approximate the bias in MTE and LATE estimates that would arise from using judge assignment as an instrument in my context. Since there is no ultimate outcome in my setting, this is essentially a calibration exercise. I follow Klein and use the same MTE from his empirical example,

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

where I pick $\rho = 0.5$. I assume that the true MTE is with respect to the refugee factor; $v = 1 - F_r(r_i)$. Appendix Figure A4 plots this MTE.

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

$$\frac{1}{2}\sigma_p^2 \frac{\partial^2 m(p)}{\partial p^2} + \frac{1}{2}\frac{\partial \sigma_p^2}{\partial p}\frac{\partial m(p)}{\partial p} \tag{14}$$

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}.

The larger the judicial errors, the more area in the latent variable space is used in the average. Figure A5 plots the results. As expected, the bias is worse in areas of the MTE with higher curvature. However, integrating over the distribution of the instrument, the LATE bias is 7.6%, which is relatively small.

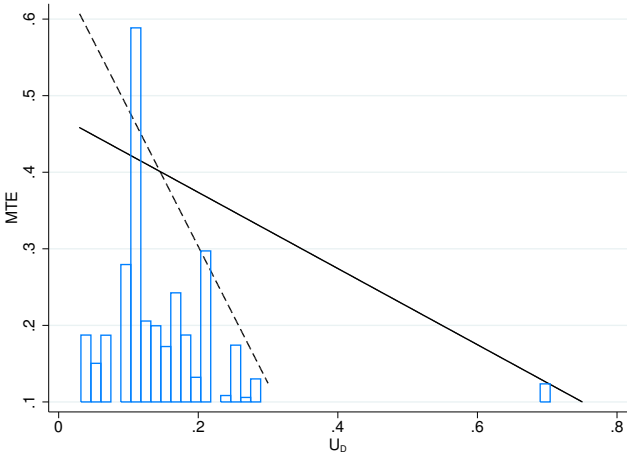
Some caution is needed in interpreting this number. First, MTE functions with more slope and curvature will imply higher bias, so this result is dependent on the context (the same MTE with $\rho = 2$ implies a bias of about 20%). Second, the estimation of the underlying quality factor r and the judicial errors $\widetilde{\varepsilon}_{ijt}$ are predicated on the first and second stage measuring the same quality. This is a reasonable assumption — the first-stage judge is tasked with determining whether the claimant would have an arguable case in the second round — but to the extent that it is violated, the estimates of the variation in the errors relative to the quality factor are inflated. This inflation would also make the estimated MTE bias larger than the true bias. Third, because decisions are not published as precedent, compared to judges in most criminal courts, FC justices may have much less information about what their colleagues would do for a given case (and therefore more variation

in what they themselves do). This would mean that this is fundamentally a court that makes a lot of errors, and that these results are not indicative of bias in other contexts.

The overall message of this section is one of cautious optimism; although judicial errors have a substantial effect on who is granted refugee status, those errors are local enough that the bias in judge-instrument designs is limited as long as the curvature of the MTE is not too high.

1.7 Appendix Figures

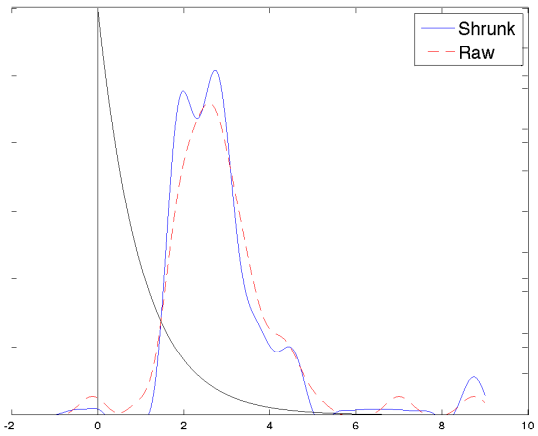
Figure A1: MTE of second-round approval on first-round approval judge



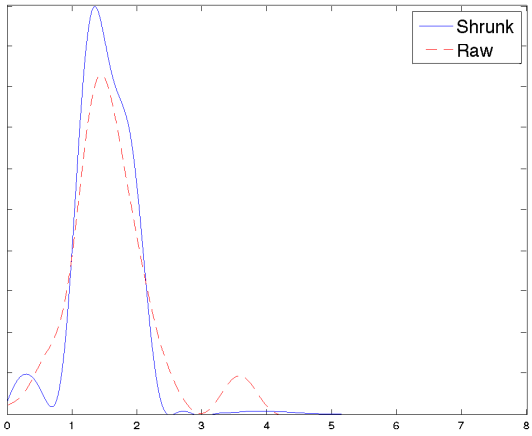
The black line represents the MTE of first-round approval on second-round approval (implicitly, no one who is rejected in the first round is approved in the second round). Estimation is from regressing a second-round approval on a second-order polynomial in the judge-level first-round approval means,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

Figure A2: Distribution of judge coefficients for flexible model

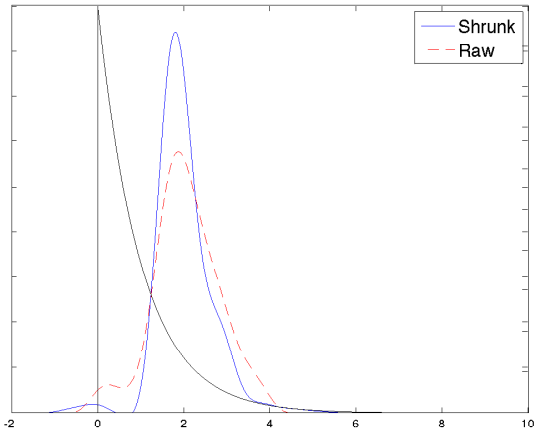
(a) Threshold γ_1, first round



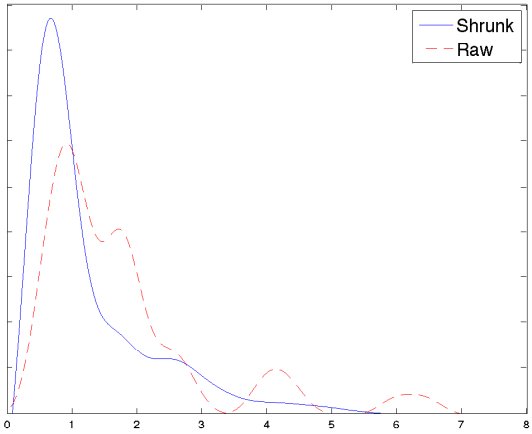
(b) Observational error σ_1, first round



(c) Threshold γ_2, second round



(d) Observational error σ_2, second round

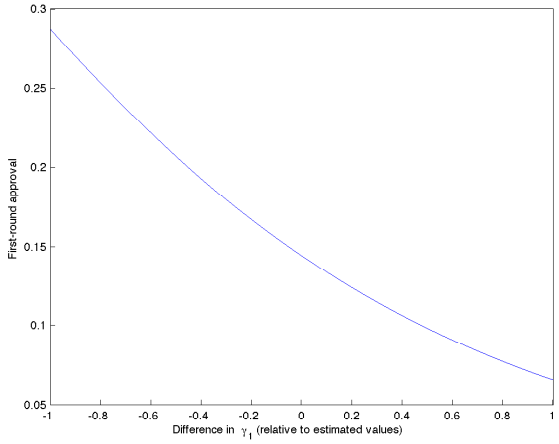


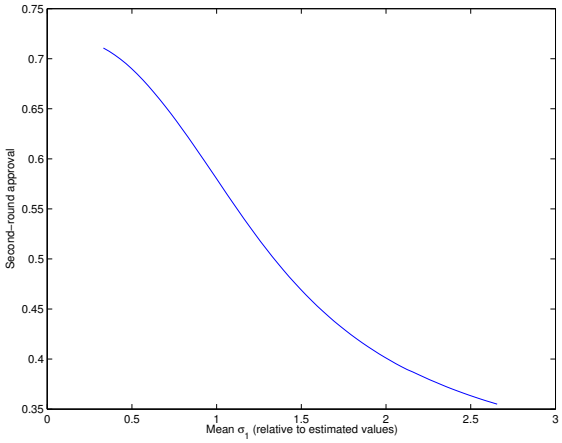
Each panel contains the density of the raw and shrunken estimates of the judge thresholds γ_1 and γ_2, and the errors σ_1 and σ_2. Black line is density of case quality r. Shrunken estimates recovered via deconvolution of estimates and accounting for heterogeneous standard errors (Delaigle and Meister, 2008).

Figure A3: Identification for model with flexible σ_1 and σ_2

(a) Mean threshold γ_1, first round

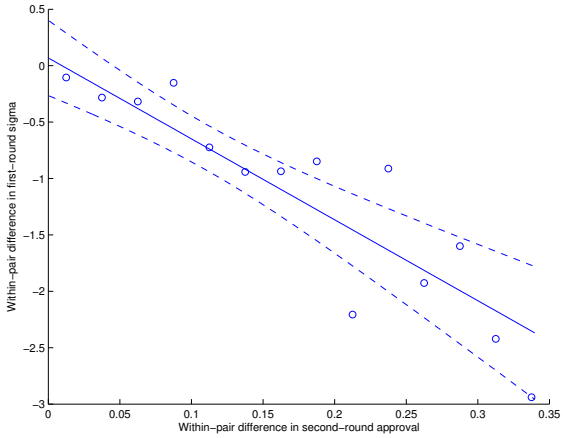
(b) Mean error σ_1, first round

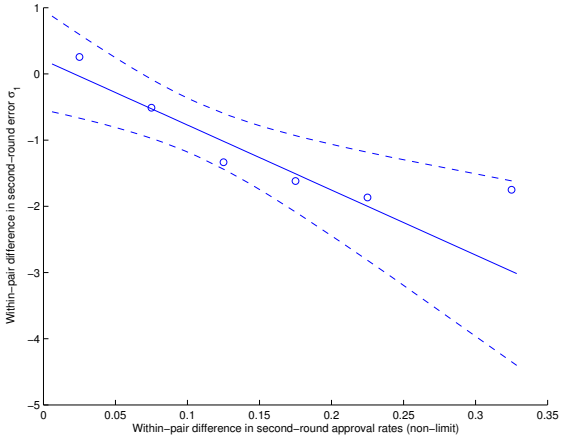




(c) Judge difference in σ_1, first round

(d) Judge difference in σ_2, second round





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 error σ_1. In Panel D, I match pairs of second-round judges with similar approval rates conditional on first-round approval by 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. See Section 2.3.1 for more details.

Figure A4: MTE used in calibration exercise

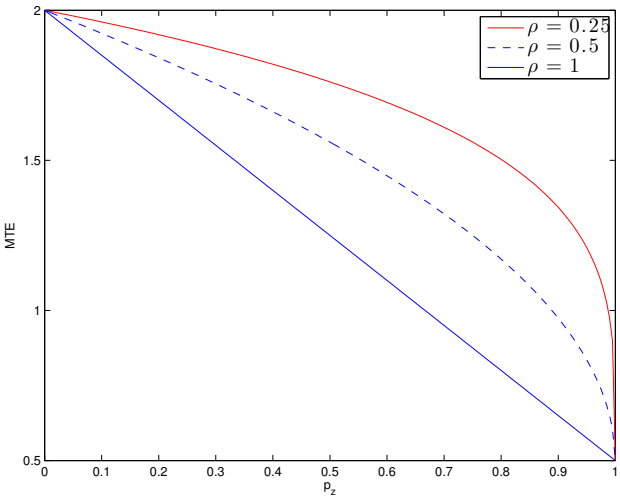
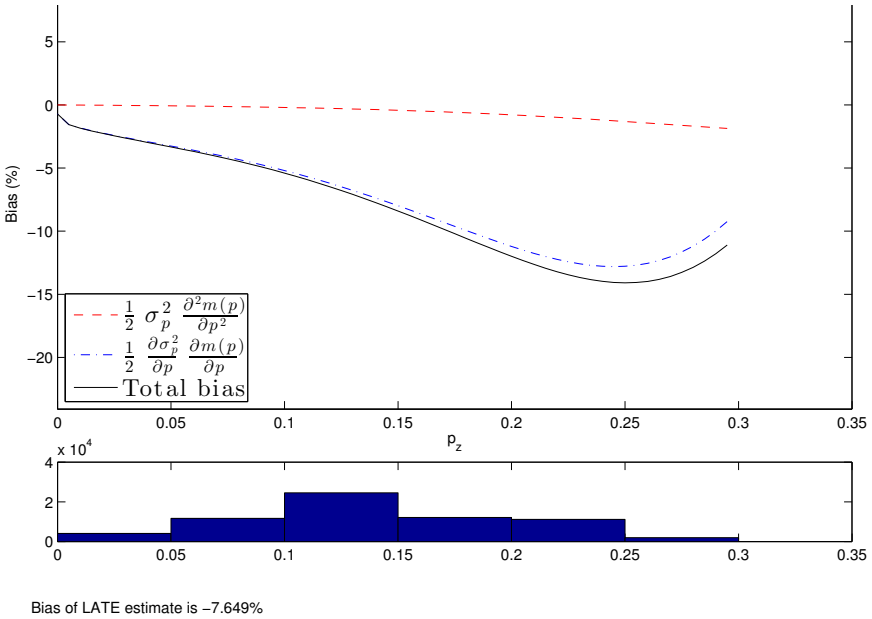


Figure A5: Estimated bias of MTE



Bias of LATE estimate is −7.649%

Estimated bias of MTE when marginal effect is $m(v) = 5 + 1.5 * (1 - v)^{\rho}$, $\rho = 0.5$. Calculations come from Klein (2010).

1.8 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 (leave cases)	-0.07	0.80	-3.45	1.53
Workload (JR cases)	-0.15	0.40	-1.01	1.35
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 Asia (=1)	0.09	0.29	0.00	1.00
Middle East (=1)	0.14	0.35	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	58604			

Table A2: Randomization

	Male Male (1)	Africa Africa (2)	Asia Asia (3)	South Asia (4)	Middle East (5)	South America (6)	Predicted approval (7)	1st-round mean approval (8)
Panel A: First round judges								
Judge mean approval	0.016	-0.077***	-0.001	0.003	-0.006	0.085**	-0.000	
	(0.015)	(0.021)	(0.027)	(0.011)	(0.010)	(0.036)	(0.002)	
F-stat	0.78	2.34	1.70	1.86	1.31	3.59	7.97	
Prob	0.89	0.00	0.00	0.00	0.05	0.00	0.00	
Observations	58455	58455	58455	58455	58455	58455	58455	
Panel B: Second round judges								
Judge mean approval	-0.031	-0.019	0.022	-0.050**	-0.027	0.058*	0.000	-0.028
	(0.029)	(0.024)	(0.032)	(0.019)	(0.026)	(0.032)	(0.001)	(0.017)
F-stat	1.02	1.11	0.97	0.95	1.29	1.21	1.84	3.06
Prob	0.44	0.26	0.54	0.58	0.06	0.13	0.00	0.00
Observations	8450	8450	8450	8450	8450	8450	8450	8450

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

Table A3: Overidentification tests for instruments

	(1)	(2)	(3)
Panel A: Coefficients for structural threshold β_s			
End-of-week hearing, second round	0.094***	0.085***	0.093***
	(0.008)	(0.008)	(0.008)
Hearing schedule over lunch, second round	0.400***	0.398***	0.395***
	(0.019)	(0.019)	(0.028)
Panel B: Coefficients for structural error ψ			
End-of-week hearing, second round		−0.032*	
		(0.018)	
Hearing schedule over lunch, second round			−0.011
			(0.053)
SD of γ_1	1.230	1.223	1.222
SD of σ_1	0.965	0.952	0.946

Reports coefficients for choice model $\mathbb{1}[r_i > X_{ijs}\beta_s + \gamma_{js} + \widetilde{\varepsilon}_{ijs}(X'_{ijs})]$, $\widetilde{\varepsilon}_{ijs}(X'_{ijs}) \sim \mathcal{N}(0, e^{\sigma_{js}+X'_{ijs}\psi})$. All models include controls for office of origination in X_{ijs}, and allow the parameters of the Pareto distribution of r_i to vary flexibly 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

	(1)	(2)	(3)
Panel A: γ_1 (mean=2.8, SD=1.31)			
Accuracy, SD	-0.147	0.067	0.115
	(0.091)	(0.108)	(0.126)
Favorability, SD		-0.300	-0.301
		(0.198)	(0.203)
Consistency, SD			-0.084
			(0.076)
Observations	174	174	174
Panel B: γ_2 (mean=2.25, SD=1.47)			
Accuracy, SD	-0.352**	0.053	-0.042
	(0.133)	(0.133)	(0.157)
Favorability, SD		-0.559**	-0.559**
		(0.222)	(0.219)
Consistency, SD			0.175*
			(0.101)
Observations	174	174	174
Panel C: σ_1 (mean=2.43, SD=2.28)			
Accuracy, SD	0.155	0.063	0.171
	(0.109)	(0.117)	(0.139)
Favorability, SD		0.125	0.141
		(0.132)	(0.129)
Consistency, SD			-0.266**
			(0.112)
Observations	174	174	174

Estimated with Hanushek (1974) correction for estimated dependent variable. All models include respondent fixed effects. Standard errors clustered at the judge level in parentheses. ** $p < 0.05$, *** $p < 0.01$.

Table A6: Approval rate for judges before and after reform

	Approval rate			Approval, year residualized		
	(1)	(2)	(3)	(4)	(5)	(6)
After reform (=1)	0.0224	0.0539	0.0531	0.0247	0.0737	0.0739
	(0.0244)	(0.0535)	(0.0558)	(0.0235)	(0.0533)	(0.0557)
Liberal appointee (=1)			-0.00567			0.00149
			(0.0216)			(0.0219)
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: Judge behavior and experience, linear year control

	(1)	(2)	(3)	(4)	(5)	(6)
	Coefficients in σ_1					
Experience	−0.019					
	(0.012)					
Experience (1st round)		−0.004				
		(0.124)				
Experience (2nd round)		0.035				
		(0.165)				
Experience > 1 year			−0.352		−0.334***	
			(0.224)		(0.070)	
Experience > 5 years					−0.052	
					(0.068)	
Experience > 10 years					−0.291***	
					(0.028)	
Experience > 1 year (1st round)				−0.706***		−0.715*
				(0.051)		(0.365)
Experience > 5 years (1st round)						−0.070
						(0.056)
Experience > 10 years (1st round)						−0.377***
						(0.057)
Experience > 1 year (2nd round)				0.257		0.125
				(0.184)		(0.120)
Experience > 5 years (2nd round)						0.140*
						(0.078)
Experience > 10 years (2nd round)						−0.105
						(0.083)
SD of γ_1	1.065	0.915	0.670	0.656	0.751	0.675
SD of σ_1	0.687	0.309	0.698	0.807	0.756	0.954
Linear year control	Yes	Yes	Yes	Yes	Yes	Yes

Reports coefficients for selection model $\mathbb{1}[r_i > X_{ijs}\beta_s + \gamma_{js} + \widetilde{\varepsilon}_{ijs}(X'_{ijs})]$, $\widetilde{\varepsilon}_{ijs}(X'_{ijs}) \sim \mathcal{N}(0, e^{\sigma_{js} + X'_{ijs}\psi})$. All models include controls for office and time/date of decision in β, and allow the parameters of the Pareto distribution of r_i to vary flexibly 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.