Measuring Examiner Consistency and Skill: Evidence from Refugee Decisions

Samuel Norris*

December 5, 2021

Abstract

Inconsistency—the extent to which different judges, doctors, and disability examiners make different decisions on identical cases—is a key measure of decision-making quality, but is difficult to study since one usually observes only a single decision per case. In this paper, I introduce new tools to measure inconsistency and study its causes in settings where one randomly-assigned judge makes each decision. First, a nonparametric approach reveals that inconsistency is high among Canadian refugee judges. 14% of cases are approved, but the average successful claimant would not be approved by another randomly-selected judge. Second, I use a novel structural model to show that inconsistency is driven by differential judge skill; this skill improves with experience, declines with workload, and is higher for independently-appointed judges. 29% of denied claimants would qualify as refugees if their case was reconsidered, suggesting serious deficiencies in the refugee adjudication system.

Keywords: partial identification, inconsistency, judges, examiner-assignment designs, refugees

*University of British Columbia. sam.norris@ubc.ca. I thank Arjada Bardhi, Lori Beaman, Gideon Bornstein, Kerwin Charles, Michael Frakes, Ezra Friedman, Jon Guryan, Seema Jayachandran, Cynthia Kinnan, Lizzie Krasner, Thomas Lemieux, Jens Ludwig, Justin McCrary, Laia Navarro-Sola, Aviv Nevo, Matt Notowidigdo, Matt Pecenco, Krishna Pendakur, Marit Rehavi, James Rendell, Brian Rendell, Jesse Shapiro, and Jeff Weaver for their useful thoughts and comments. Aaron Dewitt introduced me to the judicial review system for refugee claims. Andrew Baumberg, Catherine Dauvergne, Lori Hausegger, Sean Rehaag and several anonymous judges and law clerks offered valuable insight into the Federal Court. Daniela Santos-Cardenas provided excellent research assistance. I gratefully acknowledge the Social Sciences and Humanities Research Council of Canada for financial support through its Doctoral Fellowship Awards.

Certainty and fairness are key goals in the criminal justice system and for many administrative agencies. Inconsistency in decision-making—identical cases being treated differently by different decision-makers—violates these ideals, and is usually seen as morally undesirable (Rawls, 1972). It also generates costly uncertainty, distorts investment, and reduces welfare (Craswell and Calfee, 1986; Gennaioli et al., 2017). As a result, administrative agencies have added costly additional stages of decision-making to try to catch initial mistakes; the Social Security Administration allows up to four appeals after an initial rejection. Other agencies have developed internal training programs to improve consistency, while many jurisdictions have introduced sentencing guidelines in criminal cases to try to improve uniformity of punishment (USPTO, 2019; Adelman and Deitrich, 2008).

Despite a broad consensus on the importance of consistency, organizations' success at achieving it has proved extremely difficult to assess. This is because consistency is a measure of how decisions on the same case would be made by different people, rather than a directly observable characteristic of a single decision. Since we usually observe only one decision per case, measuring consistency and understanding how to improve it requires overcoming the fundamental problem of causal inference. It is therefore an area in which econometric tools can yield novel insights.

In this paper I introduce new tools to measure inconsistency, or the share of cases on which two decision-makers would make a different decision. I focus on settings where the treatment is binary and that satisfy the exogeneity and exclusion assumptions familiar from examiner-assignment IV. As a result, my tools are applicable in institutions involving a wide range of decision-makers: patent and Social Security examiners, bankruptcy and criminal judges, doctors, child welfare investigators, and environmental inspectors.

My analysis consists of two main parts. First, I show how to estimate sharp, informative bounds on inconsistency using only information on decisions and (possibly multiple) subsequent outcomes. Second, I develop a structural model of judge preferences and skill to explain the causes and consequences of inconsistency. The model reveals that much of the inconsistency in my setting comes from judges' differential accuracy, or ability to distinguish between high- and low-quality cases. While the model is only partially identified, it allows me to estimate informative bounds on the effects of various policies on accuracy, revealing concrete ways that policymakers could reduce inconsistency.

I use these tools to study refugee decisions. These decisions are among the most consequential ever made by government officials, with claimants' lives potentially at risk if they are improperly denied refugee status and deported back to their country of origin. Despite these life-or-death stakes, refugee judges in many countries are afforded a high degree of discretion (Ramji-Nogales et al., 2007). While this may be necessary given the complicated nature of many of the cases, it puts these institutions at particular risk of making inconsistent decisions.

My empirical setting is the Canadian Federal Court, where claimants who have been denied

refugee status by administrative decision-makers can file an appeal. The court decides whether the initial denial was improper in a two-stage sequential process; the second round occurs only if the first-round judge approves the case. I use the second-round decisions as outcomes, and focus on measuring inconsistency and accuracy among the first-round judges. The relatively large number of second-round outcomes—one for each second-round judge—makes for particularly informative bounds, and the rich data allow me to study how judge behavior changes with experience, workload, and the method through which judges are chosen.

My approach builds on previous work studying inconsistency. In one influential early paper, Partridge and Eldridge (1974) presented a number of judges with identical vignettes, allowing the direct observation of inconsistency.[1] However, concerns about the external validity of this approach have led researchers towards methods that can be applied on real-world data.[2] Fischman (2013) shows that when cases are randomly assigned to judges, the level of inconsistency is nonparametrically bounded by the cross-judge difference in average treatment rates. If one judge approves 50% of patent applications, and another judge 40%, they must disagree on at least 10% of applications. However, this bound is often uninformative: two judges who each incarcerate half of defendants might agree on all potential decisions, or none of them.

I show that subsequent outcomes can be used to substantially tighten the Fischman (2013) bounds, which provide no lower bound on inconsistency for pairs of judges with the same approval rate. In my setting, the subsequent outcomes are the second-round decisions. These decisions are informative because if two first-round judges were perfectly consistent—in other words, they would approve the exact same cases—then the cases that each sends to the second round should have the same average outcome. In contrast, the larger the difference in subsequent outcomes, the higher the implied inconsistency.

My approach is particularly informative when there are multiple subsequent outcomes. Since each outcome might differentially align with the preferences of the first-round judges, each outcome provides its own bound on inconsistency. In this paper, I take the decisions by different *second-round* quasi-randomly assigned judges as the set of subsequent outcomes. I find that because the second-round judges differ in how they make decisions, taking the intersection of these outcome-specific bounds substantially tightens the overall bound on first-round inconsistency.

To ensure that the bounds sharply characterize inconsistency, I rely on a linear programming approach (Balke and Pearl, 1997; Mogstad et al., 2018). The primitive in my analysis is f, the joint distribution of judges' potential decisions in the first and second round. The observed data—judge-specific approval rates in the first round, and conditional judge-pair approval rates in the

[1] Similarly, Hanna and Linden (2012) provide multiple teachers with identical exams. While their focus is on measuring discrimination over observable characteristics such as caste, they also find substantial variation in grades that different teachers assign to the same exam. However, this method is expensive to implement in many real-world institutions, and only possible with cases based solely on written material.

[2] For example, audits are infeasible in settings with interactions between the individual and the decision-maker.

second round—are linear combinations of the components of f, meaning that the data rule out certain values of f. Since inconsistency is also a function of f, the data may therefore rule out some values of inconsistency. I estimate the upper (lower) bound by maximizing (minimizing) average inconsistency as a function of f, restricting attention to values of f that generate the observed data. While these bounds are guaranteed to be sharp, they can be computationally infeasible when the number of judges or outcomes is large. This is the case in my setting. I instead estimate the bounds on a coarser partition of the joint distribution of judge decisions, and find that this approach still results in informative bounds.

Applying my method to the Federal Court, I find very high levels of inconsistency. Although judges approve only 14.7% of cases, I find that the average pair of judges disagrees on between 15.4 and 29.3% of cases. This means that for the average pair of judges, fewer than half of approved cases would have been approved by both judges. Such a high level of inconsistency is particularly surprising given the judges' relative specialization—refugee cases make up about 70% of their caseload—and suggests that inconsistency may be a problem in other institutions as well.

This paper also relates to a recent literature focusing on understanding the degree and consequences of variation in decision-makers' skill. Chan et al. (2021) shows that radiologists vary in their ability to detect pneumonia from chest x-rays, and that because doctors want to avoid Type I errors more than Type II errors, this ability is systematically correlated with their treatment propensity. Arnold et al. (2020) focuses on racial bias in bail decisions, and finds that judges have a lower ability to assess risk for black versus white defendants.

Similarly to these papers, I also consider a generalized index structure where decision-makers observe the true quality of a case, plus an additive error coming from a decision-maker-specific distribution. While the model would be point-identified under strict parametric assumptions—such as assuming that quality and judge observational errors were jointly normal, as in Chan et al. (2021)—I instead look within a flexible class of possible distributions. This means that consistency and other objects of interest are only partially identified, although the bounds tend to be informative.

I first use the index model to study inconsistency, which I bound to between 22.1 and 22.2%. To examine the causes of such high levels of inconsistency, I turn to studying judge-level *accuracy*, the difference in approval likelihood between the highest- and lowest-quality cases. A higher level of accuracy (holding approval rates fixed) corresponds to less disagreement, and so understanding the determinants of accuracy tells the policymaker how she can also improve consistency.

At the Federal Court, I find that accuracy looks akin to a measure of skill: it improves with experience, is lower in times of high workload, and is correlated with survey measures of judge quality.[3] The workload effect is driven entirely by judges with less than five years of experience; more

[3]While not all institutions have the ability to implement the techniques in this paper, many more can run a survey. The correlation between survey and model estimates of judge accuracy suggests that this could be a valuable way to assess reforms.

senior judges' accuracy is unaffected by workload. Most strikingly, accuracy improved dramatically after a reform that required that all candidates be vetted by independent legal experts.

My method also speaks to the effectiveness of the Federal Court in accomplishing its own goals. By law, the court is required to grant first-round approval to claimants who could make an "arguable case" in the second round. The court currently does so for only 14.7% of cases, of which 44% win second-round approval. However, my model reveals that between 29.5 and 30.2% of *all* cases would be successful in the second round, consistent with first-round judges having a relatively low ability to observe case quality. This also suggests that the court is improperly failing to approve many cases that meet its own standards.

While capacity constraints might preclude the court from considering all cases in the second round, different assignment policies could still improve the court's ability to identify claimants who have been improperly denied refugee status. In particular, variation in accuracy means that the court could update its judge assignment process to dramatically increase the number of successful claimants. Even under worst-case bounds on how different assignment rules would affect the quality of approved cases, I find that the court could increase the number of cases approved in both rounds by over 80% while maintaining its current implicit standards.

Finally, I study whether Canada is meeting its obligation to grant refugee status to all legitimate claimants. While the standard at the Federal Court is whether the government's initial decision to deny refugee status was reasonable—not whether the claimant actually qualifies as a refugee—approval by both judges means that the case is reassessed for refugee status on the merits. I use my model of judge accuracy to selection-correct the share of successful claimants who are ultimately granted refugee status into an estimate of the share of *all* claimants who would qualify as refugees under reappraisal. This number is shockingly high, at 29.3%, suggesting that the government's refugee adjudication system—despite the life-or-death stakes—is also rife with inconsistencies.[4]

1 Institutional background and data

I study refugee claims in Canda between 1995 and 2012, focusing on initially-denied claimants who have filed for judicial review of the decision at the Federal Court. The initial decisions on whether the claimant qualifies as a refugee are made by the executive branch, and the Federal Court process is to decide only if they are eligible to be reconsidered by the government for refugee status.

I begin by describing the government's decision-making process in enough detail to characterize the denied claimants who appeal to the Federal Court, then describe the Federal Court itself. Figure 1 provides a graphical depiction of the entire process.

[4] I also find that accurate judges are more likely to approve claimants who would qualify as legitimate refugees upon government re-appraisal, suggesting very large returns to improved accuracy.

1.1 Initial decisions by the Immigration and Refugee Board

Refugee decisions are made by Members of the Immigration and Refugee Board (IRB), who evaluate whether the claimant has a "well-founded fear of persecution for reasons of race, religion, nationality, membership in a particular social group or political opinion" (United Nations, 1967). Claims are non-randomly assigned to Members with expertise in either the claimant's country of origin or the reason for the claim. About 50% of claimants are approved as refugees; for those who are rejected, the Federal Court is the only avenue of appeal.[5]

In recent years there has been some concern that the cross-Member differences in approval rates are too large to be explained by possible differences in case composition (Rehaag, 2012). This is particularly troubling because it suggests that some claimants who reasonably meet the refugee standard are initially denied status, putting them in great personal risk if they are deported to their country of origin. I will return to this issue in Section 4.7, when I use my model to estimate the share of all denied claimants who would be granted refugee status if reappraised by the IRB.

1.2 Federal Court responsibilities and protocol

The Federal Court has jurisdiction over certain issues related to the federal government, with about 70% of their caseload devoted to refugee appeals. The scope of these appeals is limited. Judges must show deference to administrative decisions, and so decide only whether the initial IRB decision was reasonable, not whether it was correct (Rehaag, 2012).[6] In practice, this means that judges almost never gather new evidence, but instead review documents from the original decision as well as legal arguments submitted by each side.[7]

Success at the Federal Court requires approval by two consecutively-assigned judges (see Figure 1). First-round judges decide whether a claimant has an "arguable case" to make to a yet-to-be-determined second-round judge, who then decides whether the initial IRB decision was reasonable. Throughout the paper, I sometimes refer success in each of these stages as first- and second-round approval, respectively.

The first-round judge makes her decision after reviewing written records from the IRB and arguments written by each side. If the judge decides against the claimant, the claim is rejected and they are usually deported. If she grants *leave*, the case goes to the second round. Regardless of her decision, the first-round judge does not provide a written justification. Thus, although all judges regularly serve in both the first and second rounds, it is still difficult for any judge to directly study

[5] Nearly 65% of denied claimants file an appeal; many of those who do not gain status in another way, such as through marriage.

[6] An unreasonable decision is one where "there is no line of analysis within the given reasons that could reasonably lead the tribunal from the evidence before it to the conclusion."

[7] New evidence cannot be about the merits of the case. It must be about the government decision-making process, such as evidence that the IRB Member had made a racially prejudiced statement.

how their colleagues make decisions. This may contribute to high levels of inconsistency.

The second-round judge reads the same briefs and written records, then holds a hearing to question the lawyers for each side. The name of the first-round judge is not immediately available, and the judges typically don't obtain this information.[8] To reflect this, I model the second-round judge as unaware of the identity of the first-round judge—the first-round judge affects the second-round decision through her choice of which claimants to approve, but not through an information channel. This eliminates the possibility of multiple equilibria in the structural model.

If the second-round judge grants *judicial review*, the case is sent back to the IRB for a new decision. 38% of these claimants are eventually granted refugee status. If the claimant is denied judicial review, they are usually deported.

1.3 Data, judge characteristics, and incentives

My main data come from online Federal Court case reports.[9] The case reports contain information on the date the case was filed, the office that received the application, the names of the assigned judges, and the outcome in each round.

For each of the 53 judges active during my study period, I collected information on the date and party of appointment. Table A1 shows that 25% are female, and their dates of appointment range from 1982 to 2010. 72% of judges are Liberal appointees, the rest are Conservative. The average judge has 6.5 years of experience with a maximum of 28.

In addition to being directed to grant first-round leave to claimants who can make an arguable case in the second round the judges are encouraged to make consistent, predictable decisions (Federal Court, 2013). Given that well-performing judges can be appointed to higher courts, this seems likely to incentivize judges to approve first-round cases that they believe are likely to be approved in the second round.

The judges are instructed to make consistent, predicable decisions (Federal Court, 2013). This means that second-round judges are looking for initial IRB decisions that would be viewed as unreasonable by the legal community, with the knowledge that a small share of their denials of judicial review might be reviewed by a higher court. In turn, first-round judges search for the claimants most likely to make an arguable case in the second round, and so by backwards induction the decisions where the initial IRB denial was most unreasonable. Along with intrinsic motives, potential promotion to higher courts may also incentivize judges.

[8] It would be difficult to guess the identity of the first-round judge. In a typical year, a judge decides 15 second-round cases that have been approved by 7 unique first-round judges.

[9] Available at https://www.fct-cf.gc.ca/en/court-files-and-decisions/court-files.

1.4 Judge assignment

Judge assignment is plausibly unrelated to case and defendant characteristics in both stages. For the first round, judges are assigned to "leave duty" using a pre-set schedule. After enough cases have accumulated at one of the offices, the leave duty judge receives them all on one day, and is responsible for making the first round decisions for them. There is no review of the cases before they are assigned, and the leave duty schedule is not public. In the second round, cases are divided between judges without review of the contents and a computer program slots the cases into available hearing times.

Court officials believe that this process results in quasi-random assignment of cases to judges. If so, then measures of judge behavior should be uncorrelated with predetermined case characteristics that come from linked IRB case files. To test this, each row of Table 1 regresses a different characteristic on judge mean approval rates, conditioning on office-month fixed effects to account for temporal and spatial shocks to case strength.[10] For all but one of the 23 regressions, there is no statistically significant relationship between judge severity and characteristics, suggesting that the judge are indeed assigned quasi-randomly. Most reassuringly, a one percentage point more lenient judge is only 0.001 (SE=0.001) and 0.006 (SE=0.007) percentage points more likely to be assigned a claimant *predicted* to be approved in the first and second round, respectively.[11]

2 Measuring inconsistency

In Section 2.1 I first provide a formal definition of inconsistency, and then show how using outcomes can tighten analytic bounds on inconsistency. In Section 2.2 I introduce a different method to sharpen the bounds, and in Section 2.3 I present the results.

2.1 Analytic bounds on inconsistency

The researcher observes approval decisions $Y_s^i = \{0, 1\}$ for individual i in round s, so Y_1^i denotes the leave decision and Y_2^i the judicial review decision (except where it would be unclear, I leave the i index implicit). For any two judges j and k, define δ_{jk} as the share of cases they would disagree on. Using a potential outcomes framework, let $Y_s(j)$ be an indicator for approval under judge j in stage s. Focusing on first-round decisions, inconsistency between j and k is

$$\delta_{jk} \equiv P[Y_1(j) \neq Y_1(k)]$$

[10] The Federal Court is a national court (the judges spend time in each office), and so judges can receive cases from any office. However, the vast majority of cases are at either the Toronto or Montréal office. I combine the other offices into a single group.

[11] I estimate predicted approval by regressing approval in that round on the claimant characteristics in this table and taking the predicted value.

Inconsistency is a particularly useful measure of performance when the institutional objective cannot be expressed in terms of outcomes. Researchers cannot hope to measure whether a defendant is guilty "beyond a reasonable doubt," and so evaluating courts by the number of Type I or Type II errors is conceptually infeasible.[12] In these situations, inconsistency is an important measure of decision-making quality.

To aggregate the measures of judge-pair inconsistency into one overall measure, I focus on the following parameter:

$$\delta \equiv \sum_{j,k} w_{jk}\delta_{jk} \tag{1}$$

where w is chosen so that δ reflects the probability that two randomly selected judges disagree.

Let D_s^j indicate assignment to judge j. I suppose throughout that potential outcomes are independent of assignment, so $\big(\{Y_1(j)\},\{Y_2(j)\}\big) \perp \big(\{D_1^j\},\{D_2^j\}\big)$ where braces index over judges.[13] The researcher observes the likelihood of approval in each round, as well as the judges involved in the decisions: $E[Y_1 \mid D_1^j]$ and $E[Y_2 \mid D_1^j, Y_1{=}1, D_2^k]$. Fischman (2013) uses the first-round decisions only (in their context, there are no outcomes or second-round decisions) to provide the following lower bound on inconsistency:

$$\delta_{jk} \geq \underline{\delta}_{jk} = \max\big\{E[Y_1 \mid D_1^j] - E[Y_1 \mid D_1^k],\ E[Y_1 \mid D_1^k] - E[Y_1 \mid D_1^j]\big\}$$

which immediately implies

$$\delta \geq \sum_{j,k} w_{jk}\underline{\delta}_{jk} \tag{2}$$

This reflects the intuition that if one judge approves 40% of cases and another judge 50%, they must disagree at least 10% of the time. However, this lower bound tends to be quite low.[14] In particular, for judges with the same approval rate, we cannot rule out that they agree on all cases.

Having access to subsequent outcomes can substantially tighten the bound. Taking second-round approval by judge ℓ as the outcome of interest, in Appendix A I use Fréchet inequalities to

[12] In contrast, when the goal is known, it is possible to use more direct measures of examiner performance (Arnold et al., 2017; Chan et al., 2021). However, as I discuss in Section 3.1, the ability to accurately discern signals decreases inconsistency similarly to how it improves performance in these other settings.

[13] In other settings, where second-round court decisions are not used as the outcome, the second round assignment and potential outcomes would index over different possible outcomes. For expositional clarity in this setting, however, I use the same j indices.

[14] The upper bound on inconsistency also tends to be high; in this case it is 90%.

show that under quasi-random assignment of judges

$$\delta_{jk} \geq \underline{\delta}_{jk}(\ell) = \max\big\{E[Y_1|D_1^j]\big(1-2E[Y_2|D_1^j,Y_1{=}1,D_2^\ell]\big) - E[Y_1|D_1^k]\big(1-2E[Y_2|D_1^k,Y_1{=}1,D_2^\ell]\big),\\ E[Y_1|D_1^k]\big(1-2E[Y_2|D_1^k,Y_1{=}1,D_2^\ell]\big) - E[Y_1|D_1^j]\big(1-2E[Y_2|D_1^j,Y_1{=}1,D_2^\ell]\big)\big\} \tag{3}$$

This bound is informative even when both judges have the same approval rate $E[Y_1|D_1]$, at $2E[Y_1|D_1] \times \left|E[Y_2|D_1^k,Y_1{=}1,D_2^\ell] - E[Y_2|D_1^j,Y_1{=}1,D_2^\ell]\right|$. Thus, combining $\underline{\delta}_{jk}$ and $\underline{\delta}_{jk}(\ell)$ for all available ℓ will provide a weakly more informative bound than just using $\underline{\delta}_{jk}$. These judge-pair bounds can be combined into an overall bound on average inconsistency:

$$\delta \geq \sum_{j,k} w_{jk} \max\big\{\underline{\delta}_{jk}, \max_\ell\{\underline{\delta}_{jk}(\ell)\}\big\} \tag{4}$$

By construction, this lower bound on inconsistency is weakly larger than the one constructed using only first-round outcomes in (2). However, it may not be sharp in the sense of containing all information from the conditional means. This is because it does not impose that second-round judges behave consistently across their interactions with different first-round judges—in other words, $\underline{\delta}_{jk}(\ell)$ is calculated separately from $\underline{\delta}_{km}(\ell)$. In the following section, I show how an alternative approach can be used to construct sharp bounds.

2.2 Sharp bounds on inconsistency

An indirect way to characterize the possible degree of cross-judge inconsistency is to express both the target parameter (inconsistency) and the observed moments in terms of the same underlying primitive. A possible level of inconsistency is consistent with the data only if there is a value of the primitive that could generate both that degree of inconsistency and the observed data.

The natural primitive in this case is the joint distribution of judges' first and second round decisions $\big(\{Y_1(j)\},\{Y_2(j)\}\big)$, $f : \{0,1\}^{2J} \to [0,1]$. Since f is a distribution, we know that

$$\sum_m f_m = 1 \tag{5}$$

$$f_m \geq 0 \quad \forall m \tag{6}$$

Furthermore, both the conditional probabilities of approval in each round and average inconsistency (Equation (1)) can be expressed in terms only of f:

$$E\big[Y_1 \mid D_1^j\big] = \sum_m Y_1^{i(m)}(j) f_m \tag{7}$$

$$E\left[Y_2 \mid D_1^j, Y_1{=}1, D_2^k\right] E\left[Y_1 \mid D_1^j\right] = \sum_m Y_2^{i(m)}(k) Y_1^{i(m)}(j) f_m \tag{8}$$

$$\delta(f) = \sum_{j,k} w_{jk} \sum_m \left(Y_1^{i(m)}(j) + Y_1^{i(m)}(k) - 2Y_1^{i(m)}(j) Y_1^{i(m)}(k) \right) f_m$$

where f_m is an element of f and $i(m)$ maps the element of the partition of judge decisions into an individual (note that the decisions on different individuals in the same element are identical by construction).[15]

Since each of (5)-(8) place restrictions on f, they implicitly restrict the possible values of average inconsistency. To formalize this argument, it will be useful to define $\mathcal{F}_{NP}$, the set of f that is consistent with the above restrictions:

$$\mathcal{F}_{NP} \equiv \{f \in \mathcal{F} : (5), (6), (7), \text{ and } (8) \text{ are all satisfied}\}$$

For any possible value of average inconsistency δ, it is consistent with the observed data only if there is an f such that $\delta = \delta(f)$. The union of all such δ constitutes the identified set:

$$\delta^\star_{NP} \equiv \{\delta \in \mathbb{R} : \delta = \delta(f) \text{ for some } f \in \mathcal{F}_{NP}\}$$

Suppose that $\mathcal{F}_{NP}$ is nonempty. Since the restrictions in (5)-(8) are linear, $\mathcal{F}_{NP}$ is convex. By continuity of $\delta(f)$, this implies that the identified set $\delta^\star_{NP}$ is an interval (Mogstad et al., 2018). Indeed, since $\delta(f)$ is linear, $\delta^\star_{NP} \in [\underline{\delta}^\star_{NP}, \overline{\delta}^\star_{NP}]$, where these upper and lower bounds can be calculated by solving two linear programs:

$$\underline{\delta}^\star_{NP} = \min_{f \in \mathcal{F}_{NP}} \delta(f) \qquad \text{and} \qquad \overline{\delta}^\star_{NP} = \max_{f \in \mathcal{F}_{NP}} \delta(f) \tag{9}$$

These bounds implicitly include those in (2) and (4), since the right hand side of each can be expressed as a linear combination of the elements of f. However, since the approach in (9) includes more information—and in particular, it imposes that each conditional mean is generated by the same joint distribution f—they will in general be tighter. In fact, they are by construction sharp in the sense of incorporating all information from the assumptions and conditional mean approval rates in (5)-(8).

Solving the optimization problems in (9), however, turns out to be computationally infeasible. This is because f has 2^{2J} elements, which is on the order of 10^{30} in my application.[16] In Appendix B,

[15] These moments reflect that in my empirical setting, outcomes (second-round decisions) are observed only conditional on approval in the first round. However, they can be easily modified to settings where outcomes are observed unconditionally.

[16] For comparison, benchmarks for linear solvers typically have 10^6 variables (Mittelmann, 2018). Using a regression

I show how to coarsen the partition of the judge decision space so that the optimization problem is feasible but the bounds are still informative.

2.3 Empirical bounds on inconsistency

In this section, I first present descriptive statistics on judge decisions. These statistics suggest that the bounds in both (2) and (4) will be informative. I then present the bounds on inconsistency estimated under both the analytic approaches in Section 2.1 and the more general approach in Section 2.2.

Reduced-form evidence on inconsistency

There is substantial variation in approval rates across judges. Panel A of Figure 2 shows a histogram of first-round approval rates. While the average approval rate is low, at 14%, four judges approved fewer than 5% of cases, and one judge approved 70% (the next highest rate is 28%).[17] As is clear from (2), any variation in first-round approval rates implies a non-zero bound on inconsistency.

If two first-round judges never disagreed, they would have the same approval rate over randomly-assigned cases. However, (3) has an even strong prediction: the subsequent outcomes of the claimants they approve should also be the same. Panel B of Figure 2 tests whether judges meet this stronger test for consistency, plotting the likelihood of second-round approval for the claimants approved by each first-round judge against that judge's first-round approval rate. For precision I residualize out the second-round judge approval rate and shrink the estimates towards the grand mean via Empirical Bayes to account for small cell sizes, although the substantive results are unchanged by these decisions.

The figure shows additional evidence of inconsistency. Judges with the same first-round approval rate differ dramatically in their ability to select claimants who are subsequently approved; for judges approving about 15% of first-round applicants, subsequent approval rates range from 38% to 48%. This is incompatible with perfect consistency among those judges, and suggests that (4) will have identifying power above and beyond (2).

Estimates of average inconsistency

The first three columns of Table 2 present three estimates of average inconsistency. Each column uses consecutively more information: column (1) uses (2); column (2) uses (4); and column (3) uses the feasible version of (9).

As discussed in Section 1.4, judges are randomly assigned only conditional on office and time.

of solve time on the number of cells in f for 3 through 15 judges, I predict that even if enough memory was available, solving the full model would take 375,000 years.

[17] While the connection to inconsistency is not as clear, there is also substantial variation in second-round approval (Panel A of Figure A1).

Throughout the remainder of the paper I condition on office-month fixed effects. I do so by running a logit regression of the first-round approval on judge dummies and office-month fixed effects, and outputting the predicted values holding office-month fixed.[18,19] The estimates of inconsistency therefore pertain to the judges' average behavior, as in Chan et al. (2021). To reduce sampling variation, I use only first- and second-round judge-pair cells with at least 5 observations. Confidence intervals are constructed using the numerical bootstrap (Hong and Li, 2020) and clustered at the office-month level. See Appendix C for more details.

Column (1) of Table 2 reports bounds on average inconsistency using (2). While the average approval rate is 14.7%, inconsistency is no lower than 8.1%, and perhaps as high as 29.4%. These bounds tighten considerably with the addition of the second round moments in (4), to [0.153, 0.294]. They tighten again when using the feasible version of (9), to [0.154, 0.293] in column (3).

This is a high level of disagreement: column (3) tells us that for all possible joint distributions of judge decisions that are consistent with the data, a random pair of judges is more likely than not to disagree on the correct decision for approved cases.[20] The data do not rule out even higher levels of inconsistency; the upper bound in column (3) implies that a random pair of judges would agree on the correct decision for an approved case only 0.6% of the time.

Finally, Figure 3 plots inconsistency over time. For each year, I calculate the empirical analogs of the moments in (7) and (8) using an Epanechnikov kernel of bandwith 4 centered on the given year, and use these to estimate inconsistency via the feasible version of (9). The figure reveals that inconsistency was largely constant over 1995-2012, or perhaps slightly increased. This may be due to the slight increase in approval rates over the study period (from 12.9% to 16.5%), which can increase inconsistency.[21]

3 Structural model of judge decisions

In this section, I present a simple index model of how judges make decisions. This model has two important features. First, it is similar to other separable index models of decision-making used to study decision-making (e.g. Iaryczower and Shum, 2012; Chan et al., 2021), but relaxes the parametric assumptions on judges' private information about the case. Second, in an index model, judges' ability to discern case quality can be represented by a single scalar quality, which I call accuracy. This is useful because it allows me to study various inputs—such as experience,

[18] For the second-round moments, I run a logit regression of second-round approval on the first- and second-round judge pair and office-month fixed effects.

[19] I adjust the fixed effects by a constant chosen so that the predicted values have the correct sample mean.

[20] Taking the lower bound from column (3), the probability that the judges would both approve the claimant is $0.147 - \frac{1}{2}0.154 = 0.07$.

[21] If two judges' decisions are independent and each approves p_j cases, $\delta_{k\ell} = p_k + p_\ell - 2p_kp_\ell$ is increasing in p_k for $p_\ell < 0.5$.

workload, and method of appointment—to judge quality and average inconsistency. I also examine the important question of whether the average claimant in my data—who has been denied refugee status once by the executive branch—would qualify as a refugee under reconsideration of their case.

3.1 Index model of decision-making

In this model, all judges agree on the ranking of refugees by quality, although they do not observe it perfectly. This assumption is likely to hold in settings where there is a substantial amount of clarity about how to decide cases at a conceptual level, but less clarity as to how that standard applies to specific situations that might arise in practice. In my setting, this is likely true: the legislation governing the Federal Court is clear that the judges are to decide whether the initial denials by the executive branch were reasonable, a well-defined concept in Canadian law. However, given the relative dearth of written decisions (see Section 1.2), it is relatively difficult for judges to find precedent that applies exactly to any particular case.

I model the judges as observing case quality r_i with some additive error ε. They approve the case if quality is higher than some judge and round-specific threshold:

$$Y_s^i = \mathbb{1}[r_i - \gamma_{js} - \varepsilon_{ijs} > 0]$$

I am agnostic as to how the threshold γ_{js} is determined, and in particular about whether judges decide this threshold with awareness of their own ε distribution as in Chan et al. (2021). I assume that $r_i \perp \varepsilon_{ijs}$ for all j, s, that ε_{ijs} are independent for all j, s, and that r and ε are continuously distributed. This implies that their CDFs are continuous functions, so

$$\begin{aligned} P[Y_s^i | r_i, D_j^s] &= P[r_i - \gamma_{js} > \varepsilon_{ijs}] \\ &= F_{\varepsilon_{ijs}}(r_i - \gamma_{js}) \\ &= F_{\varepsilon_{ijs}}(F_r^{-1}(u_i) - \gamma_{js}) \end{aligned}$$

where F_x is the CDF of x and F_x^{-1} its inverse, and so u_i is the case quality percentile. Define the screening function as

$$p_{js}(u) \equiv F_{\varepsilon_{ijs}}(F_r^{-1}(u) - \gamma_{js}) \tag{10}$$

which maps u into the likelihood of approval by judge j in round s.

The screening function is upwards sloping, which means that for each judge, claimants with a higher value of u are more likely to be approved. This restriction flows from the judges' statutory goal of approving cases that can make a strong case in the second round—imposing that the screening functions are upwards-sloping imposes that cases that are more likely to be approved in the first

round are also more likely to be approved in the second.[22] However, as I show in Section 4.7, high-u_i claimants are also more likely to qualify as refugees under government reappraisal, suggesting that the judges' ordering of cases also reflects a broader measure of quality.

The judges' goal of approving cases that are likely to be approved in the second round also suggests that the ability to differentiate between high- and low-quality cases is an important measure of judge skill. I take as my measure of judge *accuracy* the difference in approval rates between cases with $u=1$ and $u=0$. This also relates to the classic Sah and Stiglitz (1986) measure of examiner skill, the slope of the screening function at a particular point. They refer to judge j as more discriminating than judge k at $u=z$ if $p'_{js}(z) > p'_{ks}(z)$. My measure of judge skill, instead of evaluating the slope at a particular point, takes the average derivative over the range of u. I refer to judge j as more accurate than judge k if $\Delta_j > \Delta_k$, where Δ is defined in the following way:

$$\Delta_j \equiv \int_0^1 p'_{j1}(u)\ du = p_{j1}(1) - p_{j1}(0) \tag{11}$$

It is natural to think that judges themselves would prefer to have a more upwards-sloping screening function. Holding approval rates fixed, the claimants approved by first-round judges with a higher Δ_j are more likely to be in turn approved in the second round—a natural way to interpret the judges goal of approving cases that can make an "arguable case" in the second round (see Section 1.2).

I study linear combinations of Δ_j, which includes averages and differences across groups. Since judge decision-making is encapsulated by the screening function, any given set of screening functions also corresponds to a measure of inconsistency. Figure A2 shows two pairs of judges with different screening functions, and their corresponding degree of inconsistency. Pairs of judges with similar screening functions have low inconsistency; pairs of judges with dissimilar screening functions have high inconsistency. Average inconsistency can also be constructed in the following way:

$$\delta = \sum_{j,k} w_{jk} \Big(\int_0^1 p_{js}(u) + p_{ks}(u) - 2p_{js}(u)p_{ks}(u)\ du \Big) \tag{12}$$

In Appendix D, I show that as a judge becomes more accurate, her likelihood of disagreement with another equally-lenient judge will decline under a single-crossing condition on the two screening functions. In other words, learning how to improve judge accuracy will typically also tell policymakers how to improve average consistency in the court. This is important for statistical power, because

[22] In settings with outcomes other than the second-round judge decisions, imposing that p_{j2} is upwards-sloping might be a stronger assumption. However, p_{j2} can still be interpreted as a monotone treatment response, and so in many settings could be approximated by a smooth, non-monotone function (Mogstad et al., 2018).

it allows me to use judge-level—rather than court-level—variation in characteristics to learn how to improve consistency.

3.2 Identification

Similarly to the nonparametric model, I partially identify the index model. I do so by first specifying a parametric approximation of the screening function. Since the data moments as well as all of the estimands I consider are functions only of p, I can then find the largest and smallest values of these quantities of interest that are consistent with the data. These largest and smallest values form bounds on the quantities of interest.

Since p is a bounded function on a closed interval, it can be well-approximated by a polynomial in u (Weierstrass, 1885). I use the set of Bernstein polynomials, which are computationally tractable and allow the researcher to impose monotonicity (Mogstad et al., 2018). I approximate $p_{js}(u)$ as

$$p_{js}(u, \theta) = \sum_{\ell=0}^{L} \theta_{js\ell} b_\ell(u)$$

where θ is the vector of coefficients, and b_ℓ is the ℓ^{th} Bernstein basis polynomial of degree L.

I consider the set of θ that satisfy the analogous restrictions to those in (5)-(8), and that produce upwards-sloping screening functions. I then maximize (minimize) the objectives in (11) and (12) within this restricted set of θ to produce upper (lower) bounds on the objects of interest.

Unlike in the nonparametric model, there is no guarantee that these bounds are sharp. This is because some of the constraints defining the set of admissable θ are nonlinear, and so the set of values of the objective consistent with this admissable θ might not be connected. This nonlinearity also means that estimation is more difficult than in the linear case. In Appendix E, I list the restrictions imposed on θ and discuss how I estimate the model.

4 Results

4.1 Average inconsistency

Columns (4)-(6) of Table 2 present bounds on average inconsistency from the parametric model, where the screening functions are approximated with polynomials of degree 3, 5, and 7, respectively. The bounds on average inconsistency are relatively similar across specifications, with the confidence interval for the most flexible specification spanning 21.7 to 22.8%.[23] Even at the lower end of this

[23] Because I allow for the moments to not exactly fit the data using the procedure outlined in Appendix E, it is not guaranteed that the bounds nest each other as the degree of polynomial grows.

range of possible inconsistency, this means that a random pair of judges would only agree on the correct decision for 26.2% of approved cases.

This is a very high level of inconsistency, and means that nearly all claimants would be granted leave by some judges but not others. For the remainder of the paper, I investigate the causes, consequences, and possible solutions to this high level of inconsistency.

4.2 Judge accuracy

The average judge exhibits a relatively high degree of accuracy, with an average Δ_j of [0.486, 0.493] when approximating the screening functions using polynomials of degree 7 (for the rest of the paper, I use $L=7$ as the baseline). The likelihood of approval ranges from [0.043, 0.044] at $u=0$ to [0.529, 0.537] at $u=1$. However, because many high-quality cases are never granted leave in the first round, it is not clear that the court is satisfying its mandate to grant first-round approval to all claimants who could make an "arguable case" in the second round. The court currently grants leave to only 14.7% of cases, of which 44% are granted judicial review (for a 6% overall judicial review rate). In contrast, the model reveals that between 29.5 and 30.2% of *all* cases would be granted judicial review if they were considered for it.[24] In other words, the court grants judicial review (second-round approval) to less than one quarter of the cases that qualify for it.

4.3 Experience and workload

The key input for optimizing judicial terms is the degree of learning; if experienced judges are substantially more accurate, then judicial churn is costly and should be avoided. Table 3 explores the returns to experience by comparing judges' accuracy when they have more versus less than five years of experience.[25] Column (1) shows that relative to a baseline accuracy of [0.323, 0.327], judges with more than 5 years of experience are [0.063, 0.071] more accurate.

Another important input to judge productivity is workload. Column (2) of Table 3 compares judges with more versus less than their median monthly workload of cases, and finds that judges are [0.039, 0.052] *less* accurate in periods of high workload. Finally, column (3) interacts workload with more versus less than 5 years of experience. Consistent with judges improving over time at managing higher workloads, there is no statistically significant difference in accuracy between high- and low-workload periods for judges with more than 5 years of experience. In contrast, inexperienced judges' productivity declines by a statistically significant [0.057, 0.094] in periods of high workload.[26]

[24] Another way to read this statistic is that the court would grant judicial review to between 27 and 27.8% of cases that are rejected for leave in the first round.

[25] I include judges who are never observed with both more and less than five years of experience in the model, but their accuracy does not enter the estimand.

[26] The difference in workload effect across experienced versus inexperienced judges is [0.070, 0.116] with a 95% CI of (0.028, 0.181).

These results have clear takeaways for policy. The court's average accuracy could be improved by increasing judge retention, reducing workload, and transferring cases from inexperienced to experienced judges.

4.4 Judge characteristics and method of selection

Another way that policymakers might improve accuracy is through the choice of which judges they appoint to the court. Table 4 shows cross-judge differences in accuracy by various characteristics. While these characteristics may be correlated with other judge characteristics—and so can't be interpreted as causal—they are suggestive of several dynamics in the judge selection process.

First, column (1) shows that male judges are substantially less accurate than their female colleagues, at [0.404, 0.411] versus [0.609, 0.613]. While these quantities speak to average accuracy, rather than marginal, they are suggestive of a higher bar for female judge candidates. Second, column (2) compares accuracy for judges appointed by the Liberal versus Conservative parties. Despite the Liberals filling the vast majority of positions (see Table A1), the Liberal judges are [0.067, 0.078] less accurate. This is consistent with the marginal Liberal judge being less accurate than the marginal Conservative judge, perhaps because of the higher number of appointments.

Another way to improve judge quality might be through changing the method of selection. The largest reform to judge selection in recent history occurred in 1988, when a new law required that the Minister of Justice select new judges only from a list of candidates approved by an independent committee. The committee consisted mostly of members of the broader legal community who were not directly appointed by the government, and assessed the candidates in terms of competence, fairness, and ethical standards (Hausegger et al., 2010).

The standards had bite—only 40% of candidates were approved by the committee—and seem to have fulfilled the legislative goal of reducing the number of judges with close ties to the ruling party.[27] An important question is whether this reform also improved judge accuracy.

Columns (3)-(5) of Table 4 compares judges appointed before versus after the policy change, restricting to judges appointed within 15, 10, and 5 years of the reform. The judges appointed after are consistently more accurate. In the tightest comparison (column (5)), judges appointed in the five years after the reform are [0.130, 0.139] more accurate than judges appointed in the five years before the reform. This is despite no change in average approval rates across the two groups, and means that reducing the discretion of politicians to pick judges viewed as unqualified by the legal community has increased consistency.

[27]Before the reform, at least 47% of appointed judges had some political involvement, ranging from financial contributions to running for office (Russell and Ziegel, 1991). Conversely, only 30% of post-reform judges made political donations in the five years before their appointment (Hausegger et al., 2010).

4.5 Identifying judge accuracy through surveys

It might not be practical for all institutions to estimate this structural model themselves. In this section, I consider the question of how else they might learn about judge accuracy. I focus on surveys, which most decision-making institutions do have the capacity to administer. To the extent that a survey can measure the same underlying primitives as the model, it could serve as a more accessible way of measuring judge quality.

In 2017, I conducted an email survey of refugee lawyers who had appeared at the Federal Court. I asked respondents to rate judges with whom they had personal experience along several dimensions: first, leniency towards claimants, and second, consistency and predictability conditional on leniency (more details about the survey are in Appendix G). Each response was on a five-point likert scale.

I am particularly interested in whether judges who are seen as consistent and predictable have higher levels of accuracy. To implement this analysis, I divide judges by whether their average rating was above or below the median in each category, and then compare the average Δ_j across the two groups. Table 5 contains the results.

There is no statistically significant relationship between surveyed favorableness towards claimants and model accuracy (column (1), 95% CI ranges from -0.041 to 0.065). However, columns (2) and (3) show that judges who were rated as either highly consistent or highly accurate have higher levels of model accuracy (by [0.092, 0.111] and [0.138, 0.151], respectively). I conclude that in settings where my structural analysis cannot be repeated, surveys of other institutional actors might be a low-cost way of evaluating judges as well as policy reforms aimed at improving judge accuracy.

4.6 Optimal allocation of judges

There is substantial variation in accuracy across judges. This suggests that alternative mechanisms of allocating judges to cases might be superior to the current system. In this section, I study how to use judge assignments to either minimize judge workload fixing total approvals, or maximize approvals fixing workload—both without negatively affecting the quality of approved cases. The details of the optimization problems used to construct these alternative allocations are in Appendix F1.

In columns (1) and (2) of Table 6 I study the problem of minimizing judge workload. In line with conversations with court officials, I assume that second-round cases (which unlike first-round decisions, require an in-person hearing and a full written decision) take 10 times as long as first-round cases. I search over the set of possible allocations of cases to pairs of first- and second-round judges, restricting attention to allocations where all cases are adjudicated, where no judge works more than in the current allocation, and where the distribution of quality u for approved claimants first-order stochastically dominates the distribution under the baseline allocation for all admissable values of

θ. Column (1) shows that workload could be reduced by up to 24%.[28] Given the improvements in accuracy resulting from lower workload that I discussed in Section 4.3, this is likely to understate the benefits of such a policy. In column (2), I impose the additional restriction that each judge works at least half as much in each round as she did in the baseline allocation, ensuring that judges continue to gain experience at all stages of the process. While this reduces the gains to reallocation, they are still about 12%.

In columns (3)-(5) of Table 6 , I study the alternative goal of maximizing the number of cases that are granted judicial review (approved in the second round), subject to the constraint that the quality of the approved cases continues to first-order stochastically dominate that of the approved cases under the baseline allocation for all admissable values of θ. In column (3), I find that the court could increase the number of cases granted judicial review by an enormous 81%. This remains true even if we require that each judge works as much as in baseline (column (4)). Finally, in column (5), I impose the restriction that each judge works at least half as much in each round as she did in the baseline allocation, and find that the number of approvals could still be increased by 73%. These findings are important because they mean that the court is missing nearly half of the cases that would qualify—by its own standards—for judicial review under a different but feasible case assignment process.

4.7 How many legitimate refugees does Canada reject?

Recall that my sample consists of refugee claimants who are appealing their denial of status by the IRB (see Section 1.1 and Figure 1). Advocates have questioned the performance of the IRB, but it has been difficult to conclusively measure the fairness of the current system (Rehaag, 2012). In this section I use my model to estimate the likelihood that the average claimant who was denied refugee status by the IRB would be granted status if their case was considered anew by a different decision-maker.[29] I do so by taking the small subset of claimants who were approved in both rounds at the Federal Court—and as a result, sent back to the IRB for a new decision—and using the model to selection-correct their approval rate to the average for the entire population of denied claimants.

Of the claimants who are approved in both rounds at the Federal Court, I observe a new IRB decision—which I denote as Y_3—for 89% of them. I assume that claimants who I do not observe are not granted status. I then assume that the average approval rate of claimants of quality u if re-evaluated by the IRB can be well-approximated by a Bernstein polynomial of degree M.[30] By searching within the admissable set of parameters that generate the observed IRB approval

[28] This would save the court approximately $5.3 million in judge salaries alone over the study period.

[29] About 35% of claimants who are denied status by the IRB do not appeal to the Federal Court. Since often this because the claimant has gained status through another method (such as marriage), I do not model this decision and take the 65% of claimants who do appeal as representative of the broader group of denied claimants.

[30] In contrast to the screening functions, I do not assume that this function is monotonic in u.

rates for the cases granted judicial review by a given pair of judges at the Federal Court (formally, $E[Y_3 \mid D_1^j=1, Y_1=1, D_2^k=1, Y_2=1]$), I construct bounds on $E[Y_3]$, the average approval rate for all cases who appeal to the Federal Court. See Appendix F2 for more details.

Table 7 contains the results. I consider approximating the average approval rate at the IRB by polynomials of degree 3, 5, and 7 (in all specifications I use the baseline specifications of judge behavior, approximating the screening function with a polynomial of degree 7). Across columns, the results are similar: approximately 29% of all claimants at the Federal Court would qualify for refugee status if reconsidered by another IRB decision-maker (see bottom row in table). In contrast, because of the requirement that they be granted judicial review before being heard again at the IRB, only 2.4% of the rejected claimants who file an appeal at the Federal Court ever receive refugee status.[31]

As I discuss in Section 1, the judges' objective is to determine whether the IRB decision-maker's denial of refugee status was "reasonable," not whether it was correct. However, an important question is whether the high-u cases that are more likely to win judicial review (be approved in both rounds) are also more likely to be granted refugee status upon re-review by the IRB. If so, this suggests that there are advantages to judge accuracy above and beyond consistency across judges: improving accuracy will also help the court more effectively identify claimants who have been improperly denied refugee status. It also suggests that the judges' ordering of cases corresponds to a broader notion of quality.

The main estimates in Table 7 show the difference in success at the IRB for claimants with $u=1$ vs claimants with $u=0$, or $E[Y_3 \mid u=1] - E[Y_3 \mid u=0]$. Consistent with the ordering of cases by u also reflecting a more general notion of quality, 42.9% of high-u claimants would qualify as refugees if they were re-evaluated by the IRB, in contrast to 0% of the low-u claimants. This means that improving judge accuracy would let the Federal Court more effectively correct government decision-makers' erroneous denials of refugee status. However, given the very low rates of success at the Federal Court over my study period, the vast majority of denied claimants who meet the governments own standard for being a refugee are never granted status.

5 Conclusion

This paper considers the *consistency* of decision-making institutions, or the average likelihood that two different decision-makers (such as judges) would disagree on the correct decision for a randomly selected case. Previous work has identified bounds on inconsistency using differences in approval rates; if one judge approves 40% of cases and another 50%, they must disagree on at least 10%

[31] Of the claimants who are rejected by the Federal Court in either of the rounds, approximately 28% would be approved by the IRB upon re-evaluation.

of cases (Fischman, 2013). I begin by showing that these bounds can be substantially tightened by considering future outcomes. My bounds are also sharp, in the sense of using all available information from the provided moments. I estimate a computationally feasible (but non-sharp) version of my estimator, and bound inconsistency to between 15.4 and 29.3% for a sample of Canadian refugee judges. Given that average approval rates are only 14.7%, this means that the average pair of judges is more likely to disagree than agree on a given case.

I next turn to understanding the causes and consequences of inconsistency. I consider an index model of decision-making, where judges observe a common signal of quality with some additive error. I focus on accuracy, or the judge-specific difference in approval rates between the highest- and lowest-quality cases. My model of decision-making is flexible, and analogously to the first part of the paper, the quantities of interest are only partially identified. Nonetheless, I estimate informative bounds on how accuracy changes with experience, workload, and the method of judge selection. I also consider the optimal allocation of judges to cases, and find a number of policy changes that could substantially improve accuracy and consistency at the court.

Finally, this paper speaks to the extent to which Canada improperly denies refugee status to qualified claimants. The cases in my sample have all been denied refugee status by the government; if approved by judges in both rounds they are reconsidered *de novo* by a new administrative decision-maker. I use the model to calculate the share of *all* cases that would qualify as refugees under re-evaluation, and find that it is a shockingly high 29.3%. Given the high stakes of refugee decisions—by definition, improperly denied claimants are sent back to countries where they are in danger—this suggests that Canada's system of administrative adjudication of refugee claims is inadequate.

Figures

Figure 1: Refugee decision-making institutions in Canada

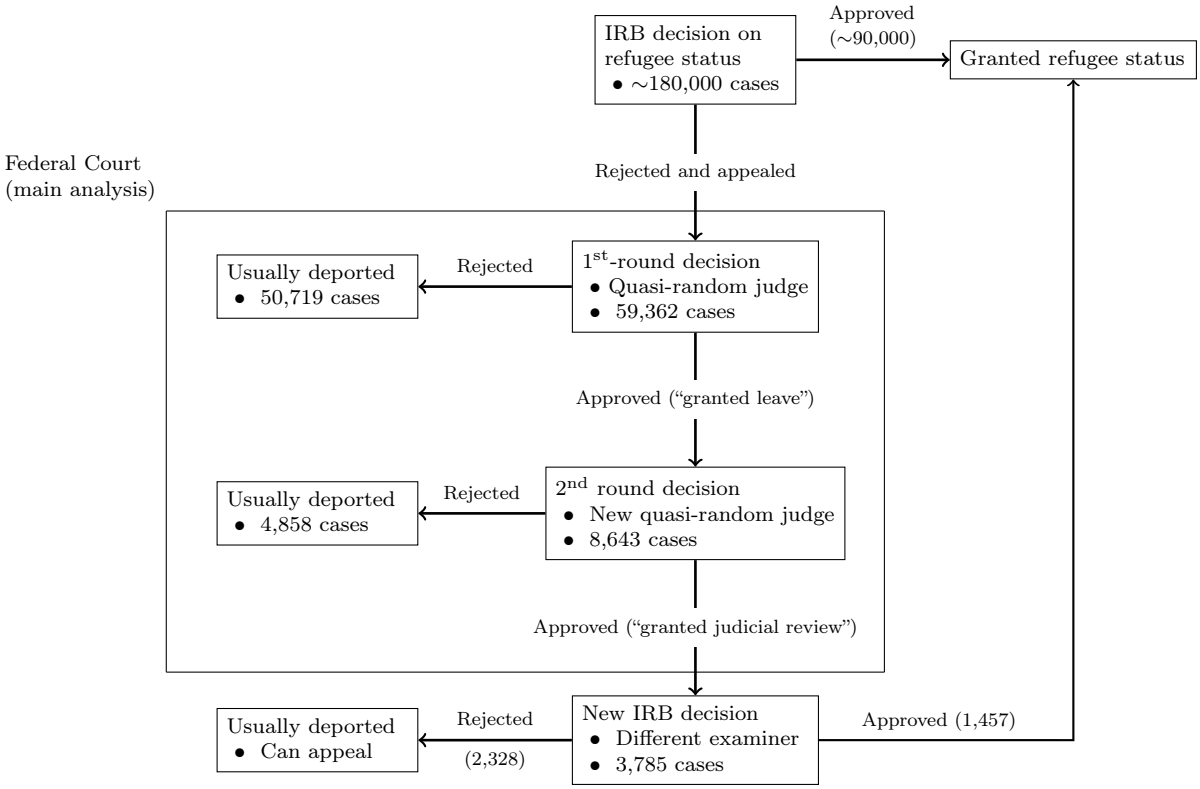
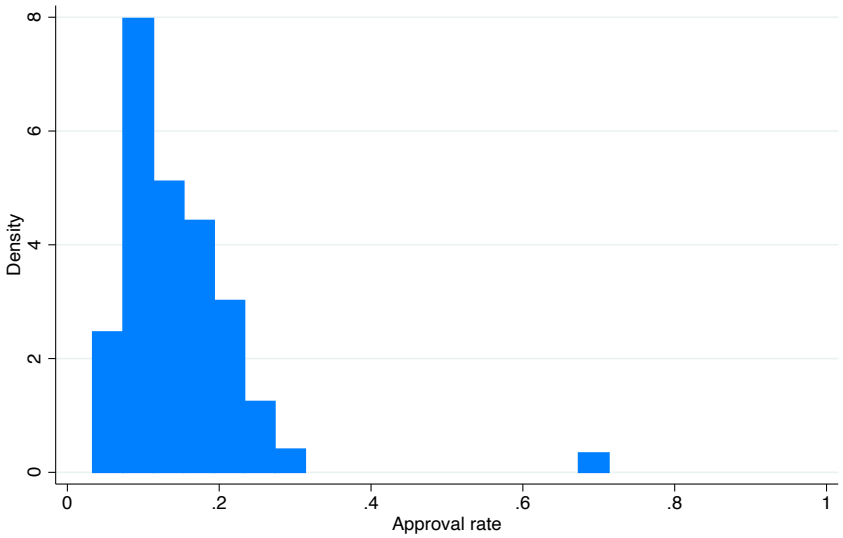


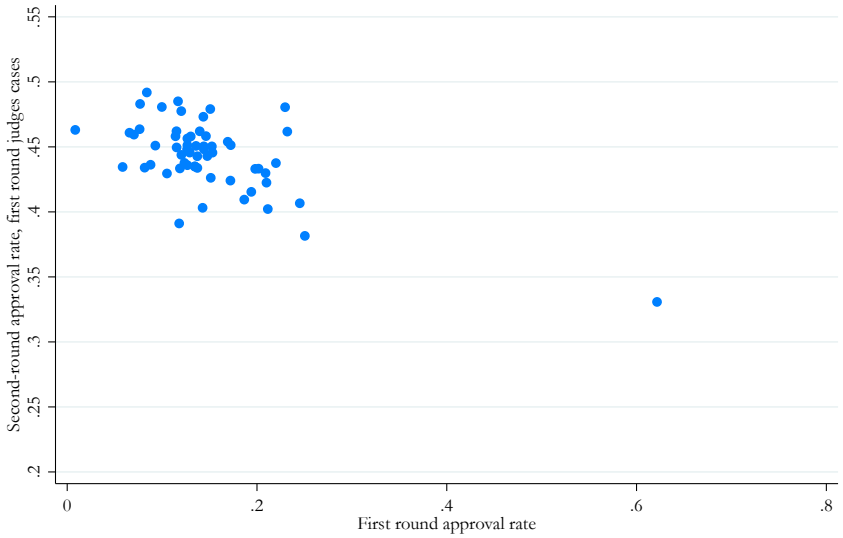
Figure represents the passage of cases through the Immigration and Refugee Board (IRB) decision on whether the claimant qualifies as a refugee, and the Federal Court's review of that decision. See institutional discussion in Section 1. I focus on the Federal Court's decisions (inside the box) for most of the paper, but discuss what we can learn about the IRB's decision-making process in Section 4.7.

Figure 2: Reduced-form judge behavior

(a) First round approval rates, by judge



(b) Subsequent approvals vs. 1st-round approval rate



Panel A contains the histogram of first-round approval rates by judge, weighted by the number of observations per judge. Panel B shows second round approval rates for the claimants approved by each first round judge plotted against that judge's first-round approval rates, with second-round judge approval rates residualized out and means shrunk via Empirical Bayes to account for measurement error from small cells. See discussion in Section 2.3.

Figure 3: Disagreement by year

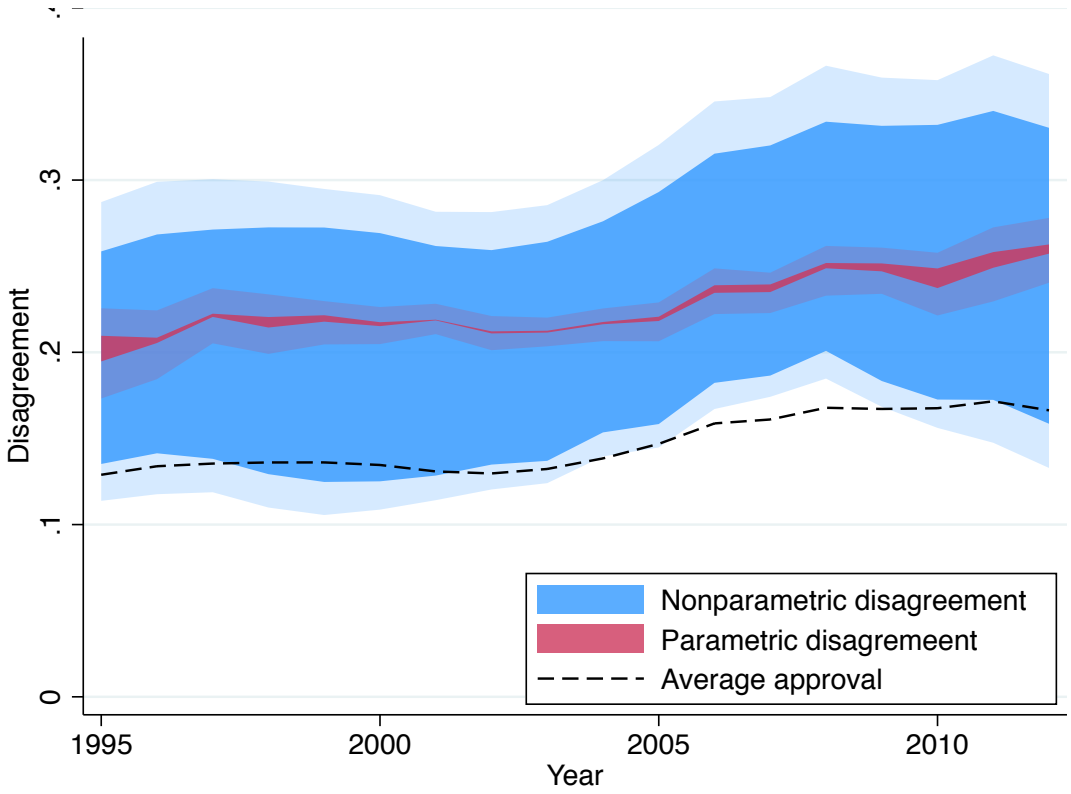


Figure reports estimated year-specific bounds on disagreement for the nonparametric model in blue, and for the parametric model in red. Nonparametric disagreement calculated using joint bounds from the feasible version of (9), and parametric disagreement using the flexible method in Section 3 with a polynomial of degree 7 to approximate the judge screening functions. For each year, I estimate disagreement for the 4 years on either side of the target year, weighted using an Epanechnikov kernel. Data moments account for office-month fixed effects estimated in logit model. Pale areas denote 95% confidence intervals, clustered at the office-month level and calculated using the numerical bootstrap (Hong and Li, 2020). Dashed line represents Epanechnikov-weighted average approval in each year. See discussion in Section 2.3.

Tables

Table 1: Placebo tests for judge assignment

Characteristic	First round		Second round	
	Mean	β	Mean	β
Predicted approval	0.144	0.001	0.434	0.006
		(0.001)		(0.007)
Male	0.650	-0.007	0.632	-0.079
		(0.028)		(0.056)
African	0.043	-0.015	0.030	0.011
		(0.012)		(0.019)
Caribbean	0.124	0.001	0.130	0.049
		(0.018)		(0.035)
European	0.144	-0.007	0.142	-0.015
		(0.022)		(0.035)
Hispanic	0.225	-0.009	0.202	-0.017
		(0.022)		(0.043)
Muslim	0.200	-0.012	0.208	-0.048
		(0.024)		(0.043)
East Asian	0.068	0.026	0.081	0.019
		(0.017)		(0.031)
South Asian	0.112	0.040**	0.114	-0.032
		(0.019)		(0.032)
Unknown origin	0.083	-0.025	0.092	0.032
		(0.017)		(0.032)
IRB officer mean approval	0.440	0.013	0.461	0.030
		(0.010)		(0.019)
1^{st}-round judge avg approval			0.198	-0.009
				(0.018)

Table shows placebo tests for judge assignment in the first and second round. For each characteristic, β comes from a regression of the characteristic on judge average approval, conditioning on office-month of assignment. Standard errors clustered by office-month in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2: Bounds on average disagreement

	Nonparametric, by method			Parametric, by degree L		
	First round	First- and second-round	Joint	3	5	7
	(1)	(2)	(3)	(4)	(5)	(6)
Average disagreement	[0.081, 0.294]	[0.153, 0.294]	[0.154, 0.293]	[0.235, 0.235]	[0.227, 0.228]	[0.221, 0.222]
	(0.079, 0.306)	(0.149, 0.306)	(0.150, 0.305)	(0.230, 0.242)	(0.224, 0.235)	(0.217, 0.228)
Average approval	0.147	0.147	0.147	0.147	0.147	0.147

Table reports estimated bounds on disagreement in square brackets. Column (1) separately calculates bounds on disagreement for each pair of first-round judges using only first-round moments, then combines them into a bound on average disagreement using (2). Column (2) separately calculates bounds for each judge pair using first- and second-round moments, then combines them into an overall bound using (4). Column (3) uses the same moments as column (2), but estimates bounds on average disagreement using a feasible version of the joint method in (9). Columns (4)-(6) measure disagreement using the flexible parametric model in Section 3, approximating the judge screening functions with polynomials of degree L. Data moments in all models account for office-month fixed effects estimated from a logit model. 95% confidence intervals in parentheses clustered at the office-month level calculated using the numerical bootstrap (Hong and Li, 2020). See discussion in Section 2.3 and Section 4.1.

Table 3: Accuracy by judge experience and workload

	(1)	(2)	(3)
More than 5 years experience	[0.063, 0.071]		
	(0.003, 0.114)		
More than median workload		[-0.052, -0.039]	
		(-0.106, -0.010)	
More than med. workload, <5 years exp.			[-0.094, -0.057]
			(-0.139, -0.032)
More than med. workload, ≥5 years exp.			[0.008, 0.042]
			(-0.016, 0.091)
Left-out mean accuracy	[0.323, 0.327]	[0.412, 0.419]	
Diff-in-diff			[0.070, 0.116]
			(0.028, 0.181)
Difference in approval rates	0.048	-0.005	
Diff. in approval rates, <5 years exp.			-0.020
Diff. in approval rates, ≥5 years exp.			0.001

Table reports within-judge differences in accuracy by experience and workload, where accuracy is the difference in approval likelihood between the consensus highest- and lowest-quality claimants. Accuracy calculated using the flexible parametric model of Section 3, with the judge screening functions approximated using 7^{th} degree polynomials. Data moments in all models account for office-month fixed effects estimated from a logit model. 95% confidence intervals in parentheses clustered at the office-month level calculated using the numerical bootstrap (Hong and Li, 2020). See discussion in Section 4.3.

Table 4: Accuracy by judge characteristics

			Appointed within years of reform		
			15	10	5
	(1)	(2)	(3)	(4)	(5)
Male	[-0.206, -0.199]				
	(-0.285, -0.163)				
Liberal		[-0.078, -0.067]			
		(-0.140, -0.027)			
Appointed after reform			[0.122, 0.154]	[0.020, 0.033]	[0.130, 0.139]
			(0.068, 0.216)	(-0.056, 0.119)	(0.029, 0.245)
Left-out mean accuracy	[0.609, 0.613]	[0.507, 0.514]	[0.363, 0.396]	[0.399, 0.404]	[0.393, 0.400]
Difference in approval rates	0.017	0.023	0.041	0.029	-0.000
Left-out mean approval	0.135	0.130	0.114	0.117	0.128

Table reports differences in accuracy across groups of judges that defined by the characteristic in the row title, where accuracy is the difference in approval rates between the consensus highest- and lowest-quality claimants. Columns (3)-(5) use same judge characteristic (appointed after reform to selection process) but restrict the sample to judges appointed within 15, 10, and 5 years of the reform occuring. Accuracy calculated using the flexible parametric model of Section 3, with the judge screening functions approximated using 7^{th} degree polynomials. Data moments in all models account for office-month fixed effects estimated from a logit model. 95% confidence intervals in parentheses clustered at the office-month level calculated using the numerical bootstrap (Hong and Li, 2020). See discussion in Section 4.4.

Table 5: Model accuracy by surveyed judge characteristics

	(1)	(2)	(3)
Above-median surveyed favorableness	[0.006, 0.011]		
	(-0.041, 0.065)		
Above-median surveyed consistency		[0.092, 0.111]	
		(0.055, 0.171)	
Above-median surveyed accuracy			[0.138, 0.151]
			(0.097, 0.210)
Left-out mean accuracy	[0.399, 0.404]	[0.398, 0.409]	[0.324, 0.341]
Difference in approval rates	0.030	-0.002	0.023
Left-out mean approval	0.132	0.150	0.144

Table reports differences in accuracy across groups of judges that scored above- versus below-median on the given question in a survey sent to lawyers who had appeared in front of the judge. Accuracy is the difference in approval rates between the consensus highest- and lowest-quality claimants, and is calculated using the flexible parametric model of Section 3 with the judge screening functions approximated using 7^{th} degree polynomials. Data moments in all models account for office-month fixed effects estimated from a logit model. 95% confidence intervals in parentheses clustered at the office-month level calculated using the numerical bootstrap (Hong and Li, 2020). See discussion in Section 4.5.

Table 6: Changes in outcomes under optimal judge allocations

	Change in workload, fixing total approvals		Change in total approvals, fixing workload		
	(1)	(2)	(3)	(4)	(5)
Percent change	-24%	-12%	81%	81%	73%
Judges cannot work more than in baseline	Yes	Yes	Yes	Yes	Yes
Substantial work in each round	No	Yes	No	No	Yes
Judges must work as much as in baseline	-	-	No	Yes	Yes

Columns (1)-(2) display the reduction in workload for the workload-minimizing allocation of judges, holding fixed the number of 2^{nd} round approvals and ensuring that the distribution of u of the approved cases under the optimal allocation first-order stochastically dominates the distribution under the baseline allocation. Columns (3)-(5) similarly show the approval-maximizing allocation, holding fixed the workload for each judge and ensuring FOSD of u for the approved cases under the optimal allocation relative to the baseline. Substantial workload means that each judge must work at least 50% as much in each round in the optimal allocation as they did in the baseline. I assume that each 2^{nd} round case takes 10 times as long to decide on as each 1^{st} round case. FOSD imposed at 100 points in the support of u. See discussion in Section 4.6.

Table 7: Differences in success at post-court hearings, highest- vs. lowest-u cases

M	3	5	7
Post-court approval rate difference	[0.320, 0.320]	[0.376, 0.376]	[0.429, 0.429]
	(0.299, 0.384)	(0.332, 0.446)	(0.378, 0.507)
Lowest-u success rate	[0.000, 0.000]	[0.000, 0.000]	[0.000, 0.000]
Average success rate	[0.297, 0.301]	[0.295, 0.295]	[0.293, 0.293]

Table reports differences in the likelihood that the highest- versus lowest-quality claimants would be approved as refugees if considered by the government. M refers to the degree of polynomial for the post-court approval likelihood; all judge screening functions use the baseline specification. Data moments in all models account for office-month fixed effects estimated from a logit model. 95% confidence intervals in parentheses clustered at the office-month level calculated using the numerical bootstrap (Hong and Li, 2020). See discussion in Section 4.7.

References

Adelman, L. and J. Deitrich (2008): "Marvin Frankel's mistakes and the need to rethink federal sentencing," *Berkeley J. Crim. L.*, 13, 239.

Arnold, D., W. Dobbie, and C. S. Yang (2017): "Racial bias in bail decisions," *The Quarterly Journal of Economics*.

Arnold, D., W. S. Dobbie, and P. Hull (2020): "Measuring racial discrimination in bail decisions," Tech. rep., National Bureau of Economic Research.

Balke, A. and J. Pearl (1997): "Bounds on treatment effects from studies with imperfect compliance," *Journal of the American Statistical Association*, 92, 1171–1176.

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

Chan, D. C., M. Gentzkow, and C. Yu (2021): "Selection with Variation in Diagnostic Skill: Evidence from Radiologists," *The Quarterly Journal of Economics*.

Craswell, R. and J. E. Calfee (1986): "Deterrence and uncertain legal standards," *JL Econ. & Org.*, 2, 279.

Federal Court (2013): *Strategic plan 2014-2019*, Government of Canada.

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

Gallichio, S. and B. Bye (1980): "Consistency of Initial Disability Decisions among and within States. Social Security Administration, Office of Research and Statistics, Staff Paper no. 39," .

Gennaioli, N., G. Ponzetto, et al. (2017): "Optimally vague contracts and the law," Tech. rep.

Hanna, R. N. and L. L. Linden (2012): "Discrimination in grading," *American Economic Journal: Economic Policy*, 4, 146–68.

Hausegger, L., T. Riddell, M. Hennigar, and E. Richez (2010): "Exploring the Links between Party and Appointment: Canadian Federal Judicial Appointments from 1989 to 2003," *Canadian Journal of Political Science*, 43, 633–659.

Hong, H. and J. Li (2020): "The numerical bootstrap," *The Annals of Statistics*, 48, 397–412.

Iaryczower, M. and M. Shum (2012): "The value of information in the court: Get it right, keep it tight," *American Economic Review*, 102, 202–37.

Kline, P. and C. Walters (2021): "Reasonable Doubt: Experimental Detection of Job-Level

Employment Discrimination," *Econometrica*, 89, 765–792.

Mittelmann, H. (2018): "Benchmark of commercial LP solvers," Tech. rep., Arizona State University.

Mogstad, M., A. Santos, and A. Torgovitsky (2018): "Using instrumental variables for inference about policy relevant treatment parameters," *Econometrica*, 86, 1589–1619.

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

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

Ramji-Nogales, J., A. I. Schoenholtz, and P. G. Schrag (2007): "Refugee roulette: Disparities in asylum adjudication," *Stan. L. Rev.*, 60, 295.

Rawls, J. (1972): *A Theory of Justice*, Harvard paperbacks, Clarendon Press.

Rehaag, S. (2012): "Judicial Review of Refugee Determinations: The Luck of the Draw?" .

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

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

United Nations (1967): "Protocol relating to the status of refugees," *Treaty Series*, 30.

USPTO (2019): *Performance and Accountabliity Report*, United States Patent and Trademark Office.

Weierstrass, K. (1885): "Über die analytische Darstellbarkeit sogenannter willkürlicher Functionen einer reellen Veränderlichen," *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin*, 2, 633–639.

Appendix for *Measuring Examiner Inconsistency and Skill*

A Analytic bounds on inconsistency

The Fischman (2013) bounds on judge inconsistency can be tightened by using additional information on subsequent outcomes. Beginning with the definition of disagreement and expanding using the law of total probability,

$$\delta_{jk} = P[Y_1(j) \neq Y_1(k) \mid Y_2(\ell)=1]P[Y_2(\ell)=1] + P[Y_1(j) \neq Y_1(k) \mid Y_2(\ell)=0]P[Y_2(\ell)=0] \tag{A1}$$

Taking the probability of disagreement from the first term on the left side of the expression, $P[Y_1(j) \neq Y_1(k) \mid Y_2(\ell)=1]$, Fréchet inequalities imply that

$$\begin{aligned}
&P[Y_1(j) \neq Y_1(k) \mid Y_2(\ell)=1] \\
=&P[Y_1(j)=1, Y_1(k)=0 \mid Y_2(\ell)=1] + P[Y_1(j)=0, Y_1(k)=1 \mid Y_2(\ell)=1] \\
\geq& \max\left\{0, P[Y_1 \mid D_1^j, Y_2(\ell)=1] - P[Y_1 | D_1^k, Y_2(\ell)=1]\right\} + \max\left\{0, P[Y_1 \mid D_1^k, Y_2(\ell)=1] - P[Y_1 \mid D_1^j, Y_2(\ell)=1]\right\} \\
=& \max\left\{P[Y_1 \mid D_1^j, Y_2(\ell)=1] - P[Y_1 \mid D_1^k, Y_2(\ell)=1], P[Y_1 \mid D_1^k, Y_2(\ell)=1] - P[Y_1 \mid D_1^j, Y_2(\ell)=1]\right\}
\end{aligned}$$

where the inequality follows from the Fréchet bound and independence of judge assignment and potential outcomes. The analogous argument can be made for $P[Y_1(j) \neq Y_1(k) \mid Y_2(\ell)=0]$. Plugging into (A1) and expanding the max operators,

$$\begin{aligned}
\delta_{jk} \geq \max\Big\{ &\Big[P[Y_1 \mid D_1^j, Y_2(\ell)=1] - P[Y_1 \mid D_1^k, Y_2(\ell)=1]\Big]P[Y_2(\ell)=1] + \Big[P[Y_1 \mid D_1^j, y_2(\ell)=0] - P[Y_1 \mid D_1^k, Y_2(\ell)=0]\Big]P[Y_2(\ell)=0], \\
&\Big[P[Y_1 \mid D_1^j, Y_2(\ell)=1] - P[Y_1 \mid D_1^k, Y_2(\ell)=1]\Big]P[Y_2(\ell)=1] + \Big[P[Y_1 \mid D_1^k, Y_2(\ell)=0] - P[Y_1 \mid D_1^j, Y_2(\ell)=0]\Big]P[Y_2(\ell)=0], \\
&\Big[P[Y_1 \mid D_1^k, Y_2(\ell)=1] - P[Y_1 \mid D_1^j, Y_2(\ell)=1]\Big]P[Y_2(\ell)=1] + \Big[P[Y_1 \mid D_1^j, Y_2(\ell)=0] - P[Y_1 \mid D_1^k, Y_2(\ell)=0]\Big]P[Y_2(\ell)=0], \\
&\Big[P[Y_1 \mid D_1^k, Y_2(\ell)=1] - P[Y_1 \mid D_1^j, Y_2(\ell)=1]\Big]P[Y_2(\ell)=1] + \Big[P[Y_1 \mid D_1^k, Y_2(\ell)=0] - P[Y_1 \mid D_1^j, Y_2(\ell)=0]\Big]P[Y_2(\ell)=0]\Big\}
\end{aligned} \tag{A2}$$

The first argument of the max function in (A2) is equal to $P[Y_1 \mid D_1^j] - P[Y_1 \mid D_1^k]$. The second argument can be re-arranged as

$$\Big[P[Y_1 \mid D_1^j, Y_2(\ell)=1] - P[Y_1 \mid D_1^k, Y_2(\ell)=1]\Big]P[Y_2(\ell)=1] + \Big[P[Y_1 \mid D_1^k, Y_2(\ell)=0] - P[Y_1 \mid D_1^j, Y_2(\ell)=0]\Big]P[Y_2(\ell)=0]$$

$$
\begin{aligned}
&= P[Y_1 \mid D_1^k] - P[Y_1 \mid D_1^j] + 2P[Y_2(\ell)=1]P[Y_1 \mid D_1^j, Y_2(\ell)=1] - 2P[Y_2(\ell)=1]P[Y_1 \mid D_1^k, Y_2(\ell)=1] \\
&= P[Y_1 \mid D_1^k] - P[Y_1 \mid D_1^j] + 2P[Y_2(\ell) \mid D_1^j, Y_1=1]\frac{P[Y_1 \mid D_1^j]}{P[Y_2(\ell)=1 \mid D_1^j]}P[Y_2(\ell)=1] - 2P[Y_2(\ell) \mid D_1^k, Y_1=1]\frac{P[Y_1 \mid D_1^k]}{P[Y_2(\ell)=1 \mid D_1^k]}P[Y_2(\ell)=1] \\
&= P[Y_1 \mid D_1^k] - P[Y_1 \mid D_1^j] + 2\Big(P[Y_2(\ell) \mid D_1^j, Y_1=1]P[Y_1 \mid D_1^j] - P[Y_2(\ell) \mid D_1^k, Y_1=1]P[Y_1 \mid D_1^k]\Big) \\
&= E[Y_1 \mid D_1^k]\Big(1 - 2E[Y_2 \mid D_1^k, Y_1=1, D_2^\ell]\Big) - E[Y_1 \mid D_1^j]\Big(1 - 2E[Y_2 \mid D_1^j, Y_1=1, D_2^\ell]\Big) \qquad \text{(A3)}
\end{aligned}
$$

where the third line follows from Bayes Rule and the fourth from the independence of $y_2(\ell)$ and first-round judge assignment. The third and fourth lines of (A2) follow analogously. Note that this bound can be constructed for each second-round judge ℓ, so we get the following bound on δ_{jk}:

$$
\begin{aligned}
\underline{\delta}_{jk}(\ell) = \max \big\{ & E[Y_1|D_1^k] - E[Y_1|D_1^k], \\
& E[Y_1|D_1^k] - E[Y_1|D_1^j], \\
& E[Y_1|D_1^j]\big(1 - 2E[Y_2|D_1^j, Y_1=1, D_2^\ell]\big) - E[Y_1|D_1^k]\big(1 - 2E[Y_2|D_1^k, Y_1=1, D_2^\ell]\big), \\
& E[Y_1|D_1^k]\Big(1 - 2E[Y_2|D_1^k, Y_1=1, D_2^\ell]\Big) - E[Y_1|D_1^j]\big(1 - 2E[Y_2|D_1^j, D_2^\ell, Y_1=1]\big)\big\}
\end{aligned}
$$

Since the first two terms are encapsulated in $\underline{\delta}_{jk}$, I include only the latter two terms in (3).

B Feasible estimation of inconsistency

The full problem in (9) is infeasible. This is because the joint distribution of first- and second-round judge decisions, f, has 2^{2J} elements. With a relatively small J, the relevant linear problem with 2^{2J} variables is manageable, and in my setting I can calculate the bounds with up to 15 judges in each round. However, calculating it with all judges in my data would be much more difficult. I estimate the model in (9) for up to 15 judges, then run a linear regression of estimation time on the number of elements. Predicting out of sample, I estimate that it would take 375,000 years to estimate the full model with all 53 judges.

Instead, I estimate a feasible version of (9) that nonetheless delivers informative bounds. The idea is to decompose δ into a sum over inconsistency within groups that each involve a smaller number of judges, estimate inconsistency within and between these groups, and combine. Since there are a smaller number of judges within each group, the relevant joint distribution of decisions has fewer elements, and so is feasible. To do so I decompose the judges into six groups, G_g for group g. For each group g, define $\mathcal{F}^g$ as the set of distributions over the partition $\left(\{Y_1(j)\}_{j\in G_g}, \{Y_2(k)\}_{k\in G_g^2}\right)$, where G_g^2 is the group of judges with a second-round case coming from some $j \in G_g$. I define the following subset of moments that pertain only to judges in that group:

$$E\left[Y_1 \mid D_1^j\right] = \sum_m Y_1^{i(m)}(j) f_m \qquad \forall j \in G_g \tag{A4}$$

$$E\left[Y_2 \mid D_1^j, Y_1{=}1, D_2^k\right] E\left[Y_1 \mid D_1^j\right] = \sum_m Y_2^{i(m)}(k) Y_1^{i(m)}(j) f_m \quad \forall j \in G_g, k \in G_g^2 \tag{A5}$$

The set of admissable distributions of joint decisions over this finer partition is

$$\mathcal{F}_{NP}^g \equiv \{f \in \mathcal{F}^g : (5), (6), (A4), \text{ and } (A5) \text{ are all satisfied}\}$$

I analogously define $\mathcal{F}_{NP}^{jk}$ as the set of distributions over the partition $\left(Y_1(j), Y_1(k), \{Y_2(\ell)\}_{\ell\in G_{jk}^2}\right)$, i.e. the partition of the first round decisions of j and k, and of any second-round judges that take either of their cases.

Then, using $g(\cdot)$ to represent a function that maps the judge index into a group, an upper bound

on average inconsistency is

$$
\begin{aligned}
\delta &\equiv \sum_{j,k} w_{jk}\delta_{jk} \\
&\leq \max_{f\in\mathcal{F}_{NP}} \sum_{j,k} w_{jk}\delta_{jk}(f) \\
&= \max_{f\in\mathcal{F}_{NP}} \sum_{g} \Big(\sum_{j,k\in G_g} w_{jk}\delta_{jk}(f) \Big) + \sum_{j,k|g(j)\neq g(k)} w_{jk}\delta_{jk}(f) \\
&\leq \sum_{g} \Big(\max_{f\in\mathcal{F}_{NP}} \sum_{j,k\in G_g} w_{jk}\delta_{jk}(f) \Big) + \sum_{j,k|g(j)\neq g(k)} \max_{f\in\mathcal{F}_{NP}} w_{jk}\delta_{jk}(f) \\
&\leq \sum_{g} \Big(\max_{f\in\mathcal{F}^g_{NP}} \sum_{j,k\in G_g} w_{jk}\delta_{jk}(f) \Big) + \sum_{j,k|g(j)\neq g(k)} \max_{f\in\mathcal{F}^{jk}_{NP}} w_{jk}\delta_{jk}(f)
\end{aligned}
\tag{A6}
$$

Here, the last line follows because for any $f \in \mathcal{F}_{NP}$, it is easy to construct a corresponding $f^* \in \mathcal{F}^g_{NP}$ that has the same value of inconsistency. Since $\mathcal{F}_{NP}$ is a refinement of $\mathcal{F}^g_{NP}$, we can define $f^*_n = \sum_{m\in n} f_m$. Since f satisfies all of (5)-(8) and all the constraints are linear, by construction f^* satisfies (5), (6), (A4), and (A5). Similarly, since δ_{jk} depends only on the potential outcomes for judges j and k, $\delta_{jk}(f^*) = \delta_{jk}(f)$.

Thus, the last line (A6) is an upper bound on inconsistency, although in general it will not be sharp. We can use this expression to estimate feasible nonparametric bounds on inconsistency, $\delta^\star_{FNP} \in [\underline{\delta}^\star_{FNP}, \overline{\delta}^\star_{FNP}]$, where $\overline{\delta}^\star_{FNP}$ is the solution to the last line of (A6) and $\underline{\delta}^\star_{FNP}$ is the solution to the analogous minimization problem. As seen in Section 2.3, this feasible problem continues to provide informative bounds on inconsistency.

C Inference

Confidence intervals throughout are computed via the numerical bootstrap (Hong and Li, 2020). Let $\phi : m \to \mathcal{R}$ represent the optimization process that maps the moments m (for example, the left hand sides of (A4) and (A5), or of (A9) and (A10)) into a bound for a particular target parameter. The goal is to construct a confidence interval for $\widehat{\theta} = \phi(\widehat{m})$, where $\widehat{m}$ are the empirical analogs of m that have been purged of court-month fixed effects using the logit procedure discussed in Section 2.3.

In each of 200 bootstrap iterations b, I draw bootstrap weights for each cluster (court-month) from an exponential distribution with mean and variance 1 (Kline and Walters, 2021). I then construct the bootstrap analogs of $\widehat{m}$, m_b^*, adjusting for covariates via the logit procedure. I then calculate $\theta_b = \phi\big(\widehat{m} + \epsilon_N\sqrt{N}(m_b^* - \widehat{m})\big)$, where $\epsilon_N = N^{-0.25}$ (alternative powers of N give similar results), and use the distribution of $\epsilon_N^{-1}(\theta_b - \widehat{\theta})/\sqrt{N}$ to estimate the 0.025-quantile of the lower bound and the 0.975 quantile of the upper bound. I use these two estimates as the lower and upper edges of the confidence interval.

D Connecting inconsistency and judge accuracy

One of the key objects of interest in this paper is inconsistency between judges. However, it is useful for interpretability and statistical power to be able to study a judge-level characteristic that affects consistency, as opposed to directly studying consistency.

In this section, I show that inconsistency and accuracy are in fact tightly linked. As judges become more accurate, holding leniency fixed, their inconsistency with another judge will decline as long as their screening functions obey single-crossing. More precisely, suppose that judge j with screening function p_{js} becomes more accurate, and her decision-making can now be described by the new screening function p_{j^*s}. These screening functions obey single-crossing if there exists some z such that $p_{js}(z) = p_{j^*s}(z)$, $p_{js}(u) \leq p_{j^*s}(u)$ for $u > z$, and $p_{js}(u) \geq p_{j^*s}(u)$ for $u < z$.

The before versus after difference in the judge's inconsistency with some other judge k is:

$$
\begin{aligned}
&\delta_{jk} - \delta_{j^*k} \\
&= \int_0^1 p_{js}(u) + p_{ks}(u) - 2p_{js}(u)p_{ks}(u)\ du - \left(\int_0^1 p_{j^*s}(u) + p_{ks}(u) - 2p_{j^*s}(u)p_{ks}(u)\ du \right) \\
&= 2\int_0^1 \big(p_{j^*s}(u) - p_{js}(u)\big)p_{ks}(u)\ du \\
&= 2\int_0^z \big(p_{j^*s}(u) - p_{js}(u)\big)p_{ks}(u)\ du + 2\int_z^1 \big(p_{j^*s}(u) - p_{js}(u)\big)p_{ks}(u)\ du \\
&\geq 2\int_0^z \big(p_{j^*s}(u) - p_{js}(u)\big)p_{ks}(z)\ du + 2\int_z^1 \big(p_{j^*s}(u) - p_{js}(u)\big)p_{ks}(z)\ du \\
&= 0
\end{aligned}
$$

where the second equality follows because average approval rates are the same for j and j^*, and the inequality follows because p_{ks} is weakly monotonically increasing. This confirms that as judges become more accurate, holding leniency fixed, their inconsistency with other judges declines.

E Structural model

In this section I describe in more detail the estimation strategy for the parametric model. The basis of the model is the judge screening function p_{js}, which maps the percentile of the claimant's quality u into the likelihood of approval by judge j in round s.

$$p_{js}(u, \theta) = \sum_{\ell=0}^{L} \theta_{js\ell} b_\ell(u)$$

where θ is the vector of coefficients, and b_ℓ is the ℓ^{th} Bernstein basis polynomial of degree L.

I impose the following theory- and data-driven restrictions on θ:

$$\theta_{js,\ell-1} \leq \theta_{js\ell} \text{ for all } j, s, \text{ and for } \ell \in \{1, \cdots, L\} \tag{A7}$$

$$\theta_{js\ell} \in [0, 1] \text{ for all } j, s, \ell \tag{A8}$$

$$P[Y_1 \mid D_1^j] = \frac{1}{L+1} \sum_{\ell=0}^{L} \theta_{j1\ell} \quad \text{for all } j \tag{A9}$$

$$P[Y_2 \mid D_2^k, Y_1 = 1, D_1^j] = \frac{L+1}{2L+1} \frac{\sum_{\ell_1=0}^{L} \sum_{\ell_2=0}^{L} \binom{L}{\ell_1} \binom{L}{\ell_2} \binom{2L}{\ell_1+\ell_2}^{-1} \theta_{k2\ell_2} \theta_{j1\ell_1}}{\sum_{\ell=0}^{L} b_\ell \theta_{j1\ell}} \quad \text{for all } j, k \tag{A10}$$

(A7) imposes monotonicity on $p_{js}(u)$. The other restrictions are the parametric analogs of (5)-(8). (A8) restricts the probability of approval for any claimant to be bounded between 0 and 1, analogously to (5) and (6). (A9) and (A10) represent conditional mean approval rates as functions of θ, similarly to (7) and (8).

Using these equations, we can define the set of admissable screening functions:

$$\Theta_P \equiv \{\theta : \text{(A7)}, \text{(A8)}, \text{(A9)}, \text{ and (A10) are all satisfied}\}$$

We are interested again in learning about a target parameter $\beta^\star$ that is the image of some function $\Gamma^\star : \theta \to \mathbb{R}$. I consider the following target parameters:

1. **Inconsistency:** δ as in (12)

2. **Linear combinations of accuracy:** For various values of the weights w_j, I study $\sum_j w_j \Delta_j$. When w is the share of observations in the sample, this corresponds to average accuracy (as

in Table 2). The weights can also be negative, allowing me to estimate differences in average accuracy across two groups (as in columns (1) and (2) of Table 3) or the difference-in-differences estimates in the bottom row of Table 3.

Then, the identified set is

$$\mathcal{B}^\star \equiv \{b \in \mathbb{R} : b = \Gamma^\star(\theta) \text{ for some } \theta \in \Theta_P\}$$

I identify bounds on $\beta^\star$ as $\beta^\star \in [\underline{\beta^\star}, \overline{\beta^\star}]$, where

$$\underline{\beta^\star} = \min_{\theta \in \Theta_P} \Gamma^\star(\theta) \qquad \text{and} \qquad \overline{\beta^\star} = \max_{\theta \in \Theta_P} \Gamma^\star(\theta)$$

Unlike in the nonparametric model, there is no guarantee that these bounds are sharp. This is because (A10) is nonlinear (as is the inconsistency objective) and so both Θ_P and $\mathcal{B}^\star$ might not be connected.

Accounting for misspecification and sampling error

To account for the possibility that Θ_P might be empty because of misspecification or sampling error, I estimate the bounds using a procedure initially developed by Mogstad et al. (2018). The procedure calculates the upper and lower bounds of the target parameter, searching over values of θ that are near to the θ that is the closest to satisfying Θ_P. I search within the set of θ that satisfy the first two conditions for Θ_P:

$$\Theta_M \equiv \{\theta : \text{(A7) and (A8) are satisfied}\}$$

Then, I define the following expression for the distance between the model predictions and the empirical moments as

$$\Gamma(\theta) = \sum_{j=1}^{J_1} \omega_j \left| y_j - \int_0^1 p_{j1}(u, \theta)\, du \right| + \sum_{j,k} \omega_{jk} \left| y_{jk} - \frac{\int p_{j1}(u,\theta) p_{k2}(u,\theta)\, du}{\int p_{j1}(u,\theta)\, du} \right| \tag{A11}$$

where ω are weights derived from the number of observations in each judge or judge-pair cell,[1] and

[1] Letting N_j index the number of first-round observations for each judge and N_{jk} the number of observations for

y_j and y_{jk} are the empirical analogs of $E[Y_1 \mid D_1^j]$ and $E[Y_2 \mid D_1^j, Y_1{=}1, D_2^k]$. The parameters that are close to satisfying Θ_P are

$$\Theta_{P_2} \equiv \left\{ \theta \in \Theta_M : \Gamma(\theta) \leq \underset{\theta' \in \Theta_M}{\operatorname{argmin}} \, \Gamma(\theta') + \kappa_n \right\}$$

where κ_n is a tuning parameter that converges to zero with sample size (Mogstad et al., 2018). In the empirical application, I pick $\kappa_n = 0.01$. Finally, I search within this set to calculate the following lower and upper bounds on the parameter of interest:

$$\underline{\widehat{\beta}}^{\star} = \min_{\theta \in \Theta_{P_2}} \Gamma^{\star}(\theta) \qquad \text{and} \qquad \overline{\widehat{\beta}}^{\star} = \max_{\theta \in \Theta_{P_2}} \Gamma^{\star}(\theta) \tag{A12}$$

each first- and second-round judge-pair, the weights are $\omega_j = N_j / (\sum_j N_j + \sum_{j,k} N_{jk})$ and $\omega_{jk} = N_{jk} / (\sum_j N_j + \sum_{j,k} N_{jk})$.

F Alternative optimization problems

F1 Optimal judge assignment

In Section 4.6 I consider the problem of optimally allocating cases to judges. While I consider two possible objectives—minimizing caseload while approving the same number of claimants, and maximizing the number of claimants approved in both rounds while fixing judge workloads—I explain only the latter in this appendix. The former follows analogously.

The policymakers problem is to choose allocation N, where N_{jk} is the number of cases that will be assigned to first-round judge j and second-round judge k. This assignment will result in N_{jk} first-round cases for judge j to adjudicate, and $N_{jk}E[Y_1 \mid D_1^j]$ second-round cases for judge k.

I take the court's current standards as given, and look at allocations N where the distribution of u for the approved claimants first-order stochastically dominates that under the baseline allocation $\overline{N}$ for all values of $\theta \in \Theta_{P_2}$. I impose this assumption for a finite number of grid points,which amounts to looking at allocations where the probability of $u > \overline{u}$ is greater than in the baseline allocation for $\overline{u} \in \overline{U} = [0.01, 0.02, \cdots, 1]$.

The problem can then be formulated as:

$$
\begin{aligned}
\max_{N} \quad & \sum_{j,k} N_{jk} P[Y_2 \mid D_j^1, Y_1{=}1, D_2^k] P[Y_1 \mid D_1^j] \\
\text{s.t.} \quad & \frac{\sum_{j,k} P[u > \overline{u} \cap Y_1 Y_2{=}1 \mid D_1^j, D_2^k, \theta] N_{jk}}{\sum_{j,k} P[Y_1 Y_2{=}1 \mid D_1^j, D_2^k, \theta] N_{jk}} \geq \frac{\sum_{j,k} P[u > \overline{u} \cap Y_1 Y_2{=}1 \mid D_1^j, D_2^k, \theta] \overline{N}_{jk}}{\sum_{j,k} P[Y_1 Y_2{=}1 \mid D_1^j, D_2^k, \theta] \overline{N}_{jk}} \qquad \forall \overline{u} \in \overline{U}, \quad \forall \theta \in \Theta_{P_2} \\
& \sum_{k} N_{jk} + \sum_{m} \rho y_m N_{mj} \leq \sum_{k} \overline{N}_{jk} + \sum_{m} \rho y_m \overline{N}_{mj} \qquad \forall j \\
& \sum_{j,k} N_{j,k} = \sum_{j,k} \overline{N}_{j,k}
\end{aligned}
\tag{A13}
$$

In the above expression, the first constraint requires that the distribution of u for approved cases in the optimal assignment satisfies first-order stochastic dominance over the distribution in the baseline assignment for all admissable parameter values. The second requires that each judge works no more in the optimal allocation than in the baseline, using ρ to denote the amount of time it takes to adjudicate a second-round case (in units of the time it takes to decide a first-round case). Finally, the third constraint requires that all cases be adjudicated in each assignment mechanism.

Since the FOSD constraint applies to all $\theta \in \Theta_{P_2}$, it is not straightforward to directly solve (A13). I instead pursue the following iterative approach. In the each step, I solve a feasible version of (A13), substituting the FOSD constraint for one that applies for only a finite number of θ:

$$\frac{\sum_{j,k} P[u > \overline{u} \cap Y_1Y_2{=}1 \mid D_2^k, D_1^j, \theta] N_{jk}}{\sum_{j,k} P[Y_1Y_2{=}1 \mid D_2^k, D_1^j, \theta] N_{jk}} \geq \frac{\sum_{j,k} P[u > \overline{u} \cap Y_1Y_2{=}1 \mid D_2^k, D_1^j, \theta] \overline{N}_{jk}}{\sum_{j,k} P[Y_1Y_2{=}1 \mid D_2^k, D_1^j, \theta] \overline{N}_{jk}} \quad \forall \overline{u} \in \overline{U}, \theta \in \Theta_O \tag{A14}$$

In the first step, I define $\Theta_O = \{\underset{\theta \in \Theta_{P_2}}{\text{argmin}}\, \Gamma(\theta)\}$, the θ that minimizes the distance between the model and empirical moments. Then, for each step b:

1. Substituting (A14) for the FOSD constraint in (A13), I solve for the step-specific optimal judge assignment N^b.

2. Given N^b, for each $\overline{u} \in \overline{U}$ find

$$\theta^{b,\overline{u}} = \underset{\theta \in \Theta_{P_2}}{\text{argmin}} \frac{\sum_{j,k} P[u > \overline{u} \cap Y_1Y_2{=}1 \mid D_2^k, D_1^j, \theta] N^b_{jk}}{\sum_{j,k} P[Y_1Y_2{=}1 \mid D_2^k, D_1^j, \theta] N^b_{jk}} - \frac{\sum_{j,k} P[u > \overline{u} \cap Y_1Y_2{=}1 \mid D_2^k, D_1^j, \theta] \overline{N}_{jk}}{\sum_{j,k} P[Y_1Y_2{=}1 \mid D_2^k, D_1^j, \theta] \overline{N}_{jk}}$$

   I then add $\theta^{b,\overline{u}}$ to Θ_O, so that after B steps we have $\Theta_O = \left\{\underset{\theta \in \Theta_{P_2}}{\text{argmin}}\, \Gamma(\theta), \{\theta^{b,\overline{u}}\}_{b \leq B, \overline{u} \in \overline{U}}\right\}$

The algorithm iterates steps 1 and 2 until $N^b = N^{b+1}$. By construction, the FOSD condition in (A13) is then satisfied, and so N^b is the allocation that maximizes overall approval subject to the constraints.

F2 Calculating the likelihood that claimants would qualify as refugees

The judges at the Federal Court are tasked with deciding which claimants for refugee status were improperly denied status by bureaucratic decision-makers at the Immigration and Refugee Board (IRB). However, their mandate is narrow: they decide whether the the IRB's decision was unreasonable, not whether it was correct. This is a high standard; an unreasonable decision is one where "there is no line of analysis within the given reasons that could reasonably lead the tribunal from the evidence before it to the conclusion." As a result, lawyers and advocates have argued that many claimants who are denied judicial review by the Federal Court (rejected in one of the rounds) would qualify as refugees if their case was reconsidered by the IRB.

Because the standards at the Federal Court are so high, only 6% of cases are granted judicial review. These cases are sent back to the IRB for a new decision on the merits of the case, nearly always by a different examiner, and 38.5% of them are granted refugee status. In this section, I show how to use the information on the cases that were approved by both judges at the Federal Court (and thus given a new hearing at the IRB) to back out the share of *all* claimants who would be granted refugee status if their case was re-heard by the IRB.

Let $Y_3 \in \{0, 1\}$ be the outcome if the case was re-adjudicated by the IRB. The object of interest is the share of claimants who would be granted refugee status if their case was re-heard by the IRB, $E[Y_3]$. I observe only $E[Y_3 \mid D_1^j, Y_1{=}1, D_2^k, Y_2{=}1]$, the probability that a case would be granted refugee status if approved by judge j in the first round and judge k in the second round.[2] However, under the assumption that $Y_3 \perp (D_1, D_2)$, both are a function of the same underlying primitive, the likelihood a claimant of quality u is granted refugee status. Thus, the data impose bounds on the possible values of $E[Y_3]$.

I parameterize this underlying primitive, $E[Y_3|u]$, using a Bernstein polynomial of degree M:

$$p_3(u, \lambda) = \sum_{m=0}^{M} \lambda_m b_m(u)$$

Next, I define the set of admissable λ, those that are close to minimizing the moments. To do this, it useful to define a function $\widetilde{\Gamma}(\theta, \lambda)$ whose image is the distance between the empirical

[2] 11% of claimants do not appear in the IRB case records despite being granted judicial review at the Federal Court; I assume they are not granted refugee status.

moments and predictions from a particular value of the parameters:

$$\widetilde{\Gamma}(\theta,\lambda)=\sum_{j,k}\widetilde{\omega}_{jk}\Big|\widetilde{y}_{jk}-\frac{2L+1}{2L+M+1}\frac{\sum_{\ell_1=0}^{L}\sum_{\ell_2=0}^{L}\sum_{m=0}^{M}\binom{L}{\ell_1}\binom{L}{\ell_2}\binom{M}{m}\binom{2L+M}{\ell_1+\ell_2+m}^{-1}\theta_{k2\ell_2}\theta_{j1\ell_1}\lambda_m}{\sum_{\ell_1=0}^{L}\sum_{\ell_2=0}^{L}\binom{L}{\ell_1}\binom{L}{\ell_2}\binom{2L}{\ell_1+\ell_2}^{-1}\theta_{k2\ell_2}\theta_{j1\ell_1}}\Big|$$

where $\widetilde{y}_{jk}$ is the empirical counterpart to $E[Y_3 \mid D_1^j, Y_1{=}1, D_2^k, Y_2{=}1]$, and the fraction is the corresponding model prediction. Analogously to the main results, to reduce sampling error I use all judge cells for which I observe at least 5 cases considered by the IRB. Then, for any given value of θ, define the admissable set as

$$\Lambda(\theta)\equiv\Big\{\lambda\in[0,1]^M:\widetilde{\Gamma}(\theta,\lambda)\ \leq\ \min_{\lambda'\in[0,1]^M}\widetilde{\Gamma}(\theta,\lambda')+\kappa_n\Big\}$$

where λ_m is constrained to be between 0 and 1 to ensure that $p_3(u,\lambda)$ is bounded between 0 and 1. Finally, the upper and lower bounds on $E[Y_3]$ can be found by solving:

$$\max/\min_{\theta\in\Theta_{P_2},\ \lambda\in\Lambda(\theta)}\quad\frac{1}{M+1}\sum_{m=0}^{M}\lambda_m \tag{A15}$$

where the right hand side of (A15) is just the model prediction for $E[Y_3]$. In Section 4.7 I also ask how the IRB differentially approves high-u versus low-u claimants. Substituting $\lambda_M-\lambda_0$ as the objective in (A15) delivers the difference in likelihood of being granted refugee status by the IRB for $u{=}1$ versus $u{=}0$ claimants.

G Survey questions

As I discuss in Section 4.5, in 2017 I fielded a survey of lawyers who had appeared at the Federal Court. The goal of the survey was to generate alternative measures of the same parameters that are identified by my structural model.

From the court records, I located the names of 931 lawyers who had appeared in front of one of the sample judges. I was able to find online contact information for 551 of them.[3] In April 2017, I contacted the lawyers and requested that they fill out an online survey on their experience with Federal Court judges. After one reminder, 64 lawyers responded for an overall response rate of 14%.[4] Table A2 compares responders to non-responders and lawyers for whom I couldn't find contact information. The main difference is that responders are more successful, with a first-round approval rate of 27% versus 19% for non-responders. Respondents are also slightly younger, with their first recorded case coming about one year later.

Each survey asked three questions on up to four judges, personalized to reflect the justices they actually had experience with. The questions were:

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

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

2. On a scale from 1 to 5, how would you rate the listed judges in terms of **consistency?** Are 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

[3] 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.

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

decided on?

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

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

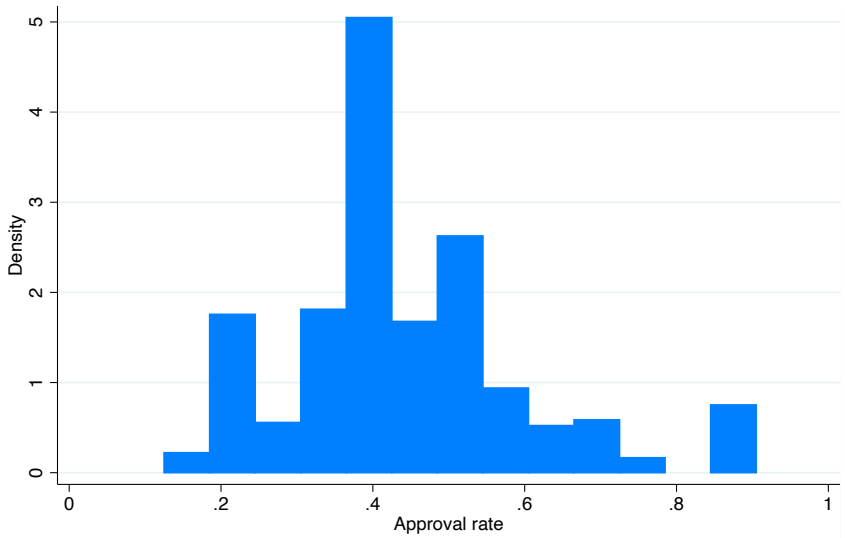
   Each question concerns one judge only, and should be answered relative to other judges. Your answer should reflect your **holistic understanding** of the judge's behavior across both leave and judicial review stages, not only the specific cases you have been involved with. Unlike the previous questions,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

I expected that the second two questions would be associated with model accuracy, but I did not have a strong prior for the relationship between model accuracy and surveyed leniency.

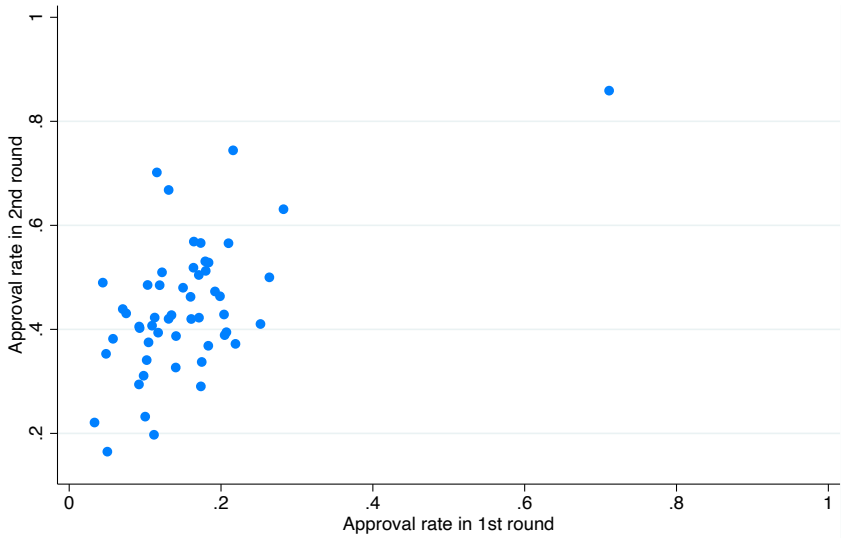
Appendix Figures

Figure A1: Second-round judge behavior

(a) Second round approval rates, by judge



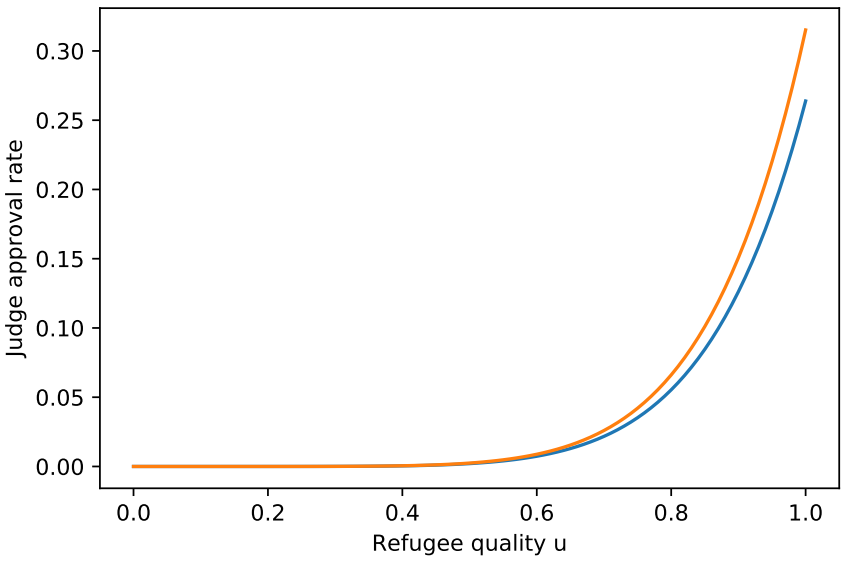
(b) First vs. second round judge approval rates



Panel A contains the histogram of second-round approval rates by judge, weighted by the number of observations per judge. Panel B contains the scatter plot of judge-level first- and second-round approval rates, shunken via Empirical Bayes to account for measurement error from small cells. The correlation is 0.57, and 0.40 without the outlier. See discussion in Section 2.3.

Figure A2: High- and low-disagreement pairs of judges

(a) Low disagreement judges



(b) High disagreement judges

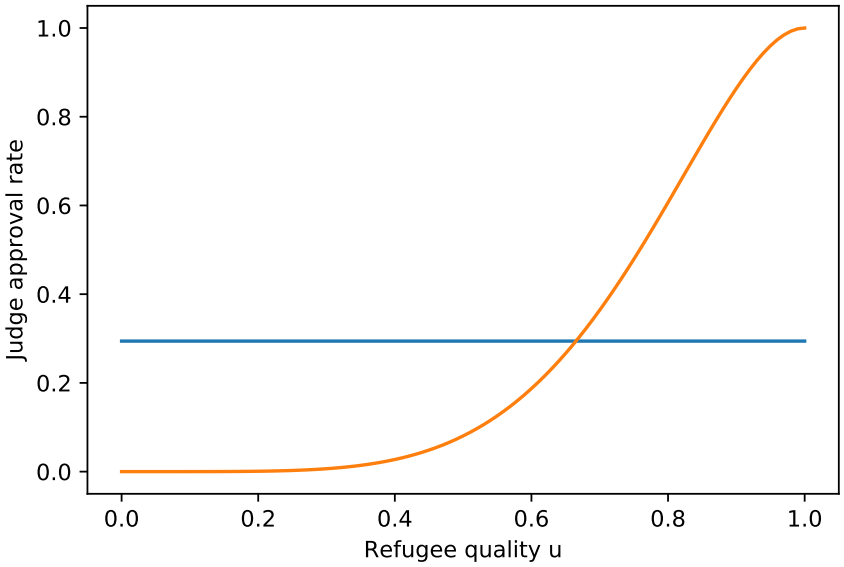


Figure shows examples of high and low disagreement pairs of judges. In Panel A, since the judges have nearly the same approval function, they disagree on only 6.1% of cases. In Panel B, in contrast, disagreement is 40.3%. See discussion in Section 3.1.

Appendix Tables

Table A1: Judge and claimant summary statistics

	Mean	SD	Min	Max
Male judge	0.74	0.44	0.00	1.00
Liberal appointee	0.72	0.45	0.00	1.00
Judge experience (years)	6.84	5.55	0.00	29.00
Judge workload	53.09	35.13	1.00	223.00
Male claimant	0.63	0.43	0.00	1.00
African claimant	0.04	0.20	0.00	1.00
Caribbean	0.12	0.33	0.00	1.00
European	0.14	0.35	0.00	1.00
Hispanic	0.23	0.42	0.00	1.00
Muslim	0.20	0.40	0.00	1.00
E Asian	0.07	0.25	0.00	1.00
S Asian	0.11	0.32	0.00	1.00
Unknown	0.08	0.28	0.00	1.00
Montréal office	0.42	0.49	0.00	1.00
Toronto office	0.50	0.50	0.00	1.00
Observations	59,362			

Workload measured at the month level and includes both first and second round cases, with second round cases counting for 10 first-round cases.

Table A2: Lawyer characteristics, survey respondents vs lawyer population

	Respondents	NR/NC	Difference
Success rate (first round)	0.27	0.19	0.074***
	[0.22]	[0.21]	(0.027)
Success rate (second round)	0.13	0.08	0.044**
	[0.16]	[0.14]	(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	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. See discussion in Appendix G. Standard deviations in square brackets and standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.