NORTHWESTERN UNIVERSITY

Three Essays in Applied Microeconomics

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Economics

By

Samuel Norris

EVANSTON, ILLINOIS

June 2018

Abstract

Three Essays in Applied Microeconomics

Samuel Norris

I study three topics in applied microeconomics. My first chapter concerns the effect of daily school start times on academic achievement in Florida. Exploiting the sharp discontinuity in school start time relative to sunrise, I track children who move between schools on either side of the time zone boundary in Florida. Consistent with children's sleep schedules becoming later after entering puberty, I find that the benefits of later start times are concentrated among girls older than 12, and boys older than 10. I conclude that reordering start times so that elementary students started first, then middle school students followed by high school students would increase math achievement for high schoolers by 0.064 standard deviations and reading scores by 0.044 standard deviations without negatively impacting younger students or affecting the overall distribution of start times in each district.

My second chapter studies how to measure the quality of *categorization* workers — judges, doctors, patent and benefits examiners — who's job is to decide whether a case fits a given criteria. I call examiners who can perfectly rank cases according to their colleagues' understanding of the standard *consistent*, and show that in institutional settings where different examiners

study the same case consistency is identified. This chapter has implications for the fundamental fairness of the justice system, by estimating the extent to which defendants sentences depend on the identity of the judge who makes the decision. It is also relevant for understanding potential biases in examiner-assignment instrumental variables research designs, which have become increasingly popular in economics over the past several years. I show that my method can quantify the extent to which the monotonicity assumption is violated, and the implications for estimating treatment effects. Using a sample of Canadian refugee judges, I find evidence of widespread differences in judges in how they rank claimants. Judges with the same overall approval rate disagree on the correct decision about 13% of the time, roughly halfway between being perfectly consistent, and flipping coins to make decisions. I find little evidence that this will induce very much bias in treatment effect estimates using examiner assignment designs, but find that marginal treatment effect estimates may be very biased.

Finally, my third chapter studies the effect of incarceration on the children of the incarcerated. Previous, correlational research has shown that the children of incarcerated individuals are very often incarcerated themselves. This may reflect either selection — the children of the incarcerated are coming from a disadvantaged background in many ways that are observable and unobservable to the researcher — or treatment — parental incarceration actually *affects* child outcomes — but pre-existing observational designs were unable to separate these two explanations. I study a natural experiment in Ohio, where the county courts randomly assign cases to judges. Judges differ in their propensity to incarcerate defendants, and I use this variation as an instrumental variable for incarceration. I find that parental incarceration *reduces* the likelihood of the child committing crimes or being incarcerated as an adult, with the reductions concentrated among children who's parent is a first time-defendant and accused of relatively

minor crimes. I find similar effects for sibling incarceration, and conclude that incarceration of a family member increases the salience of punishment for children, reducing future criminality.

Acknowledgements

I am grateful to the many people who have helped me complete this dissertation. Lori Beaman, Jon Guryan, Seema Jayachandran and Matt Notowidigdo were the best committee I could have asked for. They graciously made time to filter out my bad ideas and encourage my good ideas, give comments at my seminars and read my drafts. They have been incredible role models in terms of how to think about interesting questions, approach research, and be a constructive colleague.

I am also grateful to Krishna Pendakur, who took a chance and hired me as an undergraduate research assistant. His passion for research and constant encouragement was instrumental to me deciding to pursue a PhD.

Finally, I would like to thank Elizabeth Krasner, my parents and my siblings for their support and genuine interest in hearing about my work. My work is much better for the hours I've spent explaining it to them, and the hours they've taken to read my drafts.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Rise and Shine: The Effect of School Start Times on Academic Performance from Childhood through Puberty

1.1. Introduction

American teenagers are chronically sleep-deprived (Eaton et al., 2010). As children enter puberty, physiological changes delay the onset of sleep and make it more difficult to wake up early in the morning. By the end of middle school there is a large disconnect between physiological sleep patterns and school schedules: Hansen et al. (2005) find that students lose as much as 120 minutes of sleep per night after they start school in September, compared to the summer months when they can better control their own sleep schedules.

Sleep matters for learning and cognition. Important memory formation and consolidation processes occur overnight, as the brain replays patterns of brain activity exhibited during learning (Fogel and Smith, 2011; Maquet et al., 2000). Restricting sleep also reduces alertness and attention levels (Lufi et al., 2011; Sadeh et al., 2003), which likely affects students' ability to learn or take tests the next day. In light of these findings, the American Academy of Pediatrics recommends that adolescents wake up no earlier than 8:00 a.m. (2014). As of 2011, the median *start time* for American high schools was 8:00 a.m., suggesting that current policy may have cognitive costs for students.

Relatively little research has directly examined the effect of K-12 start times on academic performance. We study this question with a novel identification strategy that takes advantage

of the biological effect of light on sleep patterns. Sleep timing is partially regulated by sunlight exposure; holding hours of darkness constant, more sunlight in the morning (and less at night) naturally moves bedtimes earlier and increases alertness in the morning (Crowley et al., 2007). Sunlight before school — as opposed to clock start times — is therefore the correct measure of policy when comparing between schools.[1] We expect that students exposed to more sunlight will improve their academic performance, and that this effect will be stronger for pubertal children because of their delayed sleep schedules (Carskadon et al., 1997). Our empirical strategy leverages the discontinuous change in sunrise times at a time zone border, combined with the fact that school start times do not fully adjust for this difference. Using a rich administrative dataset of all public school students in Florida between 2000 and 2013, we track children as they move across the Central-Eastern time zone boundary. Treating time zone as an instrument for sunlight before school, we identify the effect of start time relative to sunrise on academic performance conditional on student fixed effects and school characteristics.

We observe children moving across the time zone boundary at all ages between eight and fifteen, which allows us to estimate the age-specific effect of school start times over a range of developmental stages. An additional hour of sunlight before school has almost no effect on math scores for pre-pubescent children, but a large and abrupt effect appears for girls at age 11 and boys at age 13. This pattern corresponds exactly to the gender-specific median age of an important pubertal transition (Campbell et al., 2012), which we take as evidence that the causal pathway is linked to the physiological changes that occur during puberty. Specifically, a one-hour delay in relative start times increases standardized math scores by 0.081 standard deviations for adolescents, but only 0.009 SDs for pre-pubertal children. In reading, an extra hour

[1] For any given school, clock start time is colinear with sunlight before school.

of sunlight before school increases scores by 0.057 SDs for adolescents and 0.061 SDs younger children. The difference between groups is not statistically significant in reading, though the adolescent estimate is more precise and can be tested as different from zero. As children move over the time zone boundary, the change in scores occurs within a year of the change in sunlight exposure and persists over time.

Later relative start times do not increase learning time for adolescents, as measured by absences. Absences are reduced by 0.869 percentage points for younger children. Differences in how absence is measured across school types (elementary, middle and high schools) may be part of the reason behind the differences in outcomes we find here. We do not observe tardiness that does not result in an absence and therefore cannot rule it out as a causal channel, but our results are consistent with improved alertness and learning capacity as a result of later start times for adolescents.

We build on the current literature in two other ways. First, we provide evidence on whether improved achievement in high-morning-sunlight areas is a result of better learning throughout the year, or merely improvements in testing performance. Using variation in test timing over the sample years, we show that testing effects are unlikely to account for the math results. They may make up a portion of the gains from later start times in reading.

Second, we address a potentially important educational policy. Although moving start times later for all students would increase academic performance at a relatively low monetary cost (Jacob and Rockoff, 2011), interference with transportation and parental work schedules is a major concern for many districts. An alternative policy is to keep the same distribution of start times, but to adjust the opening order for schools in a way that is consistent with the physiological evidence: elementary schools, middle schools, and finally high schools. We show that most districts in the Florida panhandle do not follow this optimal pattern, but that the policy

would increase math and reading scores by 0.06 and 0.04 SDs for high school students, with little negative effect for younger students. Although there may be other costs — in particular, young children might have to wait for the school bus in the dark — our paper is the first to quantify the academic benefits of this policy.

1.2. Background

1.2.1. Previous research

There have been several recent studies investigating the effect of daily start times on academic achievement, though none have examined the role that pubertal changes play in the effects. Wahlstrom et al. (1998) find that delaying school start times in Minneapolis public schools from 7:15 to 8:40 improved student sleep by 39 minutes and significantly decreased tardiness rates. Their measure of academic performance was teacher-assigned grades, where they found a positive but statistically significant effect.[2] A later paper by Hinrichs (2011) exploiting the same policy change finds no effect on ACT scores. Another approach is from Edwards (2012), who uses changes to busing schedules as a source of potentially exogeneous variation in start times. He finds evidence that delayed start times increase achievement for middle school students. The effect seems to be smaller for elementary students, but he notes that this may be a result of start times being much later for younger children in his sample. The results are not available by gender, which makes inference on the importance of puberty difficult. Finally, Carrell et al. (2011) study freshmen cadets at the United States Air Force Academy who were randomly assigned different school schedules, and who belonged to cohorts with different first-period start times. Using this random variation, they find that having a start time of 7:00 a.m. (versus

[2]Teacher-assigned grades may understate the effect of school-level interventions if teachers curve assigned grades within a given class and year.

no class in first period) decreases achievement by about 0.15 SDs in that class, and by about 0.10 SDs in subsequent classes.

1.2.2. Sunlight, sleep, and puberty

The role of sunlight in determining sleep schedules is well known. Sleep patterns are partially controlled by the circadian rhythm, which synchronizes to a 24-hour cycle using the daily variation in light and darkness (Crowley et al., 2007). In the morning, light on the outside of the eyelids suppresses production of the hormone melatonin and stimulates brain processes to increase alertness; darkness at night increases melatonin levels and feelings of tiredness (Arendt, 2000).

One of the most drastic and well-documented changes during adolescence is to the timing of sleep. As children move through puberty, nocturnal melatonin secretion is delayed several hours relative to adults and younger children (Carskadon et al., 1997, 2004). The result is that adolescent sleep patterns become more owl-like, with later bedtimes and wake times, even holding the level of darkness fixed (Carskadon et al., 1993, 2004; Crowley et al., 2007). Schools in the United States tend to begin early to accommodate after-school activities and parental work schedules, preventing adolescents from waking at their preferred later times and leading to an increasing disconnect between weekday and weekend sleep schedules during the school year (Jenni and Carskadon, 2012; Laberge et al., 2001). The result is low wakefulness and attention levels on school days (Lufi et al., 2011). More directly, sleep levels have large effects on cognitive performance (Sadeh et al., 2003; Walker and Stickgold, 2006).

Although boys and girls undergo similar sleep-related changes during adolescence, the age profile of puberty varies significantly by gender. Marshall and Tanner (1970) show that pubic

hair development begins 1.5 years earlier for girls than for boys; there is a similar gap for attainment of other developmental thresholds. This variation in age at entry into successive pubertal stages generates an important testable prediction: if physiological changes are driving the increasing importance of school start times during high school, then the size of the start time effect will co-vary with the gender-specific entry into puberty. In contrast, other changes that might make start times more relevant to achievement — e.g., the transition to a block schedule, middle-school social pressures, or changes to after-school activities — likely affect both genders at the same age.

1.3. Identification strategy

Our goal is to estimate the causal effect of school start times on academic achievement and behavioral outcomes. One approach would be to regress outcomes on start times, but because start times are chosen by the policy-maker, this approach would generate upwards-biased coefficients if better-managed schools tend to also start later in the day.[3]

Instead, our identification strategy exploits the relationship between sunlight and sleep, along with variation in sunrise time between locations. The intuition is that sleep patterns are linked partially to sunrise and sunset times, rather than clock time. This means that in terms of student sleep and alertness, the policy-relevant measure of school start time is start time relative to sunrise. For a given school, this is an unnecessary distinction: the choice of when to start classes according to the clock is equivalent to deciding when to start classes relative to sunrise. Between schools in different locations, however, a given clock start time corresponds to different relative start times. This contrast is particularly stark at a time zone boundary. Suppose that

[3]Better schools may also start earlier; for example, they may start earlier to accommodate after-school activities. This fundamental uncertainty about the direction of the bias from OLS underlines the importance of good instruments in this context.

there are two schools close together but on opposite sides of the boundary, where the sun rises at 6:00 a.m. in Central Time (CT) and 7:00 a.m. in Eastern Time (ET). If both schools begin classes at 8:00 a.m. local time, students attending the school in CT will have one more hour of sunlight before the morning bell.[4] To translate this insight into credible estimates, we track academic achievement as students move between schools on different sides of the time zone boundary. As students move from CT to ET, they are exposed to less sunlight before school, which we expect will decrease academic achievement. Conversely, a student moving from ET to CT gains sunlight before school and should see their test scores increase.

Formally, we use the time zone as an instrument for the amount of sunlight before school, which we refer to as the relative start time. We then regress academic and behavioral outcomes on instrumented relative start time to estimate the causal effect of relative start times.

The exclusion restriction in this setting is that time zone is uncorrelated with other school and student characteristics that might also affect achievement. This assumption might not be realistic in certain contexts. If, for example, we regressed achievement on instrumented time zone for the entire state of Florida, our identifying assumption would be that the only difference between schools in CT and ET relevant to student achievement is variation in relative sunrise times. Even conditional on a robust set of controls, this assumption is unlikely to hold. Instead, we include a set of student fixed effects and identify the coefficients of interest using only within-student variation. This means that variation in our instrument comes only from students who move between time zones.

[4]Children in CT will also have one less hour of sunlight after school. It is possible that this has an effect on academic outcomes, for example if less sunlight after school decreased sports participation and led to more homework time. As a policy matter, moving school start times later will always increase sunlight before school at the expense of sunlight after school; because we are interested in the effect of school start times as a policy we consider this a feature of our approach.

We relate outcomes to start times using the following functional form:

$$y_{it} = \delta_1 hours_{it} + \delta_2 hours_{it} \times \mathbb{1}[\text{puberty}] + X_{it}\beta + \gamma_i + \varepsilon_{it} \tag{1.1}$$

where y_{it} is the outcome of interest, $hours_{it}$ is the number of hours between sunrise and school start, X_{it} is a vector of controls and γ_i is an individual fixed effect. The first stage instruments for relative start time with an indicator for time zone $timezone_{it}$:

$$hours_{it} = \alpha_{11} timezone_{it} + \alpha_{12} timezone_{it} \times \mathbb{1}[\text{puberty}] + X_{it}\theta_1 + \eta_{1i} + u_{1it} \tag{1.2}$$

$$hours_{it} \times \mathbb{1}[\text{puberty}] = \alpha_{21} timezone_{it} + \alpha_{22} timezone_{it} \times \mathbb{1}[\text{puberty}] + X_{it}\theta_2 + \eta_{2i} + u_{2it} \tag{1.3}$$

where η_i are individual fixed effects. The vector X_{it} typically includes longitude, which directly affects sunrise times, as well as school-level demographic controls to proxy for school quality.

Crucially, we allow the effect of start time to vary by pubertal status. Based on the biological evidence discussed in Section 1.2.2, we expect that students' natural sleep patterns will become more out-of-sync with their school schedule as they enter puberty. We therefore expect that δ_1 in Equation 1.1 will be positive because later start times likely increase performance for children of all ages, and that δ_2 will be positive to reflect the greater benefits of later start times for adolescents.

One potential concern with this strategy is that the vast majority of cross-boundary moves are over a great distance. Long-distance moves may be inherently disruptive and therefore have an independent effect on academic outcomes. We address this concern by including in our sample students who move schools, but *not* across the time zone boundary. These students identify a set of dummies for 1, 2, and 3+ years after the move, disentangling the effect of moving from the effect of moving across a time zone boundary.

1.4. Data

1.4.1. Academic outcomes

Our data come from Florida Department of Education (FDOE) administrative records for the fifteen school years from 1998-1999 through 2012-2013 (henceforth, 1999 through 2013). We exclude alternative schools, adult education centers, and virtual academies that may have non-standard start times. Our primary outcome of interest is individual-level scores on the annual Florida Comprehensive Assessment Test (FCAT) in math and reading; this test is considered 'high stakes' for students and schools. Students took the FCAT in math in grades 5 and 8 in years 1999 through 2000, grades 3 to 10 in 2001 through 2010, and grades 3 to 8 in 2011 through 2013. They took the FCAT in reading in grades grades 4 and 8 in 1999 through 2000 and grades 3 through 10 in 2001 through 2013. Scores are standardized by year and grade at the state level for each test, with a mean of zero and a standard deviation of one. In addition to the FCAT, the data include individual-level characteristics such as race, ethnicity, gender, free- or reduced-price lunch (FRL) eligibility, and absentee rates. We use student birthdays to calculate age at the start of the school year in September.[5]

[5]The FDOE uses September 1 as the kindergarten admission cutoff.

The longitudinally-linked data allow us to follow students over time, as long as they remain within the Florida public school system. About 90% of students are matched year-to-year by social security number; the remainder are matched by name and birthday. This matching process is conducted by the FDOE and appears to contain a small number of errors caused by multiple students with similar names or birthdays. To account for this, we exclude students who move backwards more than two grades, fail and then skip a grade, have a change in birthday, are older than 15, or change gender from year-to-year. In total, these deletions amount to about 7% of the original dataset. We lose few students in the longitudinal analysis; among students who took the third grade FCAT before 2009, we observe 93% taking an FCAT the following year and over 80% taking an FCAT five years later.

We restrict the sample in two main ways to address possible threats to identification. First, we focus on the area near the time zone boundary. This reduces the likelihood that there are different economic trends on either side of the boundary, which could mean that moves in one direction were disproportionately induced by job loss. Parental job loss is often a stressor for children and may itself have a negative impact on academic achievement; this could bias our results in either direction. The area near the time zone boundary is known as the Florida panhandle, and is generally seen as distinct from the rest of the state.[6]

Second, we limit the sample to students who make a substantial move, which we define as consecutive appearances at schools further than 25 miles apart. This restriction is largely targeted at the within-time zone movers; we want to ensure that these students are subjected to something comparable to the disruptive, long-distance cross-time zone moves. The exact choice

[6]The panhandle includes the following 19 counties: Bay, Calhoun, Escambia, Franklin, Gadsden, Gulf, Holmes, Jackson, Jefferson, Lafayette, Leon, Liberty, Madison, Okaloosa, Santa Rosa, Taylor, Wakulla, Walton, and Washington. The time zone boundary approximately bisects the area.

of 25 miles as the cutoff is admittedly arbitrary; in the Online Appendix, we show that the main results are similar when using 15, 20, or 30 miles as the cutoff, or defining a move as a change in school district.

Table 4.1 displays summary statistics for third-graders in the panhandle. Note that this is a subset of our main estimation sample; we do not require that we observe a student in third grade to include them in our main analysis. However, because we intend to show that test scores are directly affected by time zone through the start time channel, observed differences in test scores for older children are not informative about baseline characteristics. The third grade summary statistics in Table 4.1 are therefore as close to baseline summary statistics as is possible with our data, although there may already be some effect of differing relative start times.

Panel A presents school-level outcomes for all students in the panhandle (Column 1); for those who move more than 25 miles (Column 2); and for those who move more than 25 miles between time zones, disaggregated by direction of move (Columns 3 and 4). Column 5 tests the difference between Columns 3 and 4. Movers come from nearly identical schools as non-movers on all dimensions. Comparing within cross-boundary movers, CT-ET movers come from fairly similar schools as ET-CT movers across most measures; two differences stand out as large and statistically significant. First, the schools in ET have a much larger percentage of black students. This occurs because most black students in our sample are from Tallahassee and its surrounding suburbs in ET. Second, the district-level third grade reading score of the cross-time zone movers' schools is 0.08 SDs higher in CT than in ET. This would be problematic for identification if it implied that underlying peer quality improves when students move from ET to CT. However, this pattern may actually be a *result* of later relative start times in CT, because these students have already been treated with four years of later relative start times in grades

K-3. In contrast, peer covariates like FRL, which are less affected by sunlight levels, are more similar between time zones. As a precautionary measure, we control for some characteristics of the peer populations with demographic share controls in our main specifications. In the Online Appendix, we show that our results are robust to the inclusion of controls for peer mean test scores.

Panel B presents individual-level characteristics. The movers are quite similar to the overall panhandle population, which bodes well for external validity. Movers are 11 percentage points more likely to be FRL relative to the non-movers, but equally likely to be black. Their test scores are slightly lower than the non-movers (0.09 and 0.08 SDs lower in math and reading, respectively), possibly reflecting stress from the upcoming move or slightly higher poverty rates among movers.

The characteristics of cross-time zone movers who begin in CT and those who begin in ET are well-balanced in terms of demographic characteristics, although the third grade math score is an insignificant 0.06 SDs lower for the CT-ET movers. The CT-ET movers also have 1 percentage point lower absentee rates than ET-CT movers.

Overall, Table 4.1 tells us that the two different types of cross-time zone movers are similar but not identical in terms of third grade characteristics and those of the schools they attend. Equality of baseline outcomes is not strictly required for our identification strategy; we make only the difference-in-differences assumption that the unobserved changes in average achievement had the students moved at a different time (or moved but not been exposed to a different relative start time) be the same for both types of mover. There are two main ways that this could be violated: if the ET-CT movers are on a different trend than the CT-ET movers, or if there are different changes in school quality over the move for different mover types.

The patterns of achievement in the years before the move provide evidence on the similarity of the underlying trend for each of the mover groups. Figure 4.1 displays pre-move trends for four types of movers — two within a time zone (CT-CT and ET-ET) and two across (CT-ET and ET-CT) — estimated from a regression of test scores on the number of years until move interacted with mover type. We include a vector of controls[7] and a fixed effect for the period preceding a move for each student. The year before the move is the excluded category. The Figure shows that the trend for each mover group is similar: in both math and reading, the test scores for each group are statistically indistinguishable from each other during the pre-move period. Time until move is also not a very strong predictor of academic achievement; for all but two of the group-time combinations, we cannot reject that there is no difference in achievement between that year and the year immediately preceding the move. This suggests that the groups are on similar underlying trajectories, and that variation in post-move outcomes can be attributed to changes in sunlight before school, rather than differential trends.

One slightly surprising finding is that math scores trend upwards for all groups in the years before the move. Long-distance moves are often a result of parental divorce or job loss, which may occur several years before the move actually takes place. Because both of these events can increase stress levels for children, it might be expected that in the absence of controls, test scores would decline leading up to a move. In the Online Appendix we confirm this intuition; in a version of the same Figure without controls we show that both math and reading scores

[7]We include all controls from our baseline regressions, which we discuss more in Section 3.5. They include age-gender dummies, longitude, and school-level demographic means (male, FRL, black, Asian, and Hispanic). The longitude and demographic coefficients are identified from small deviations in school location and school demographics in the years before the move, but have no substantive effect on the coefficients of interest. We include them for comparability with our main regressions.

unconditionally decline in the years before a move. Although we prefer the version with controls to maintain comparability with our main results, the substantive conclusion in both cases remains the same: there are no large differential trends that would threaten our identification strategy.

Another violation of our exclusion restriction would arise if school or neighborhood characteristics changed dramatically over the move. In Appendix Table A4, we present evidence that changes in these characteristics are unlikely to drive our results. Taking the year before and after each move, we regress school characteristics on a set of student-move dummies and a dummy for each of the four types of move. Relative to the schools they started in, CT-ET movers move to schools with 4.5 percentage points fewer FRL students, 14.0 percentage points more black students, and a median zip code income $5,700 higher (ET-CT movers see approximately the opposite changes). In the absence of any other intervention, this might actually raise achievement for CT-ET movers given the strong relationship between average income and school quality, when in fact we see the opposite.

1.4.2. Imputing puberty

We do not directly observe the onset of puberty, and instead use data from the National Health and Nutrition Examination Survey (NHANES) to impute developmental stage by age and gender. NHANES is a nationally representative sample of US children ages 8 to 19, and includes information on Tanner Stage, a 1-5 scale of pubertal development based on pubic hair. We use the median age of entry into Tanner Stage 3 as our cutoff for adolescence, as changes in sleep

patterns occur after the acceleration of pubertal development during Tanner Stage 3 (Campbell et al., 2012).[8]

Figure A3 in the Online Appendix displays the cumulative share of children who have reached Tanner Stage 3 by gender and age; the median age of entry occurs at 11 for girls and 13 for boys. We use these ages as the start of puberty in our analysis.

1.4.3. School start times

We define school start time as the start of the first class where learning takes place; this excludes homeroom and breakfast. Data were mostly available on school websites, and we followed up by phone with all remaining schools.

We did not collect information on historical school start times, which change with some regularity according to the school principals we spoke with while conducting the survey.[9] Given the identification strategy, our estimates will be consistent if there has been no change in the average start time for each time zone over the study period.[10] We believe that this condition is likely met: although there has been some recent discussion of school start time policy in the popular press, most of our data is from before this conversation reached the mainstream. Furthermore, the debate has never touched on whether early start times are more onerous for students with a later sunrise time.

---

[8] A second version of the Tanner Stage uses genital and breast development to demarcate stages. We use the pubic hair definition because the scale is more closely associated with pubertal changes in sleep patterns (Campbell et al., 2012), although using the alternate definition does not substantively change our main results. Using pubic hair Tanner Stage 2 or 4 changes the precision but not the direction of our results. Full results are available in the Online Appendix.

[9] This means that any attempt to estimate Equation 1.1 by OLS would result in attenuated coefficients due to measurement error on the right hand side.

[10] Under a more restrictive linear relationship between achievement and start times, we require only that there has been no change in the difference in start times between the two time zones.

School start times range from 7:00 a.m. to 9:30 a.m. local time. The average start time is 8:10 a.m., and the median is 8:00, which is similar to the national average (NCES, 2012). There is some heterogeneity with age: the median elementary school student starts school at 7:55, the median middle schooler at 8:25, and the median high schooler at 7:50. Nationwide, it is common to have high schools start earlier than the other schools in the district, so these broad patterns are not surprising.

We use NCES school location data to calculate sunrise times for each school. Combining these with our school start time data, we average the difference over the school year before the testing date to construct a measure of relative start time, measured as the number of hours between sunrise and school start times.

1.5. Results

1.5.1. First stage

Our first stage is predicated on the idea that although school start times may differ across the time zone boundary, they do not do so enough to erase the one-hour difference in sunrise times. Figure 4.3 plots the hours of sunlight before school, or relative start time, in the years before and after a move for each of the four groups of movers. We estimate each point from a regression of relative start times on time relative to move for each group as well as an individual-move fixed effect and controls for longitude and school demographics. The year before the move is normalized to be zero; we adjust the level of the coefficients with the group mean of relative start times for one year before the move.[11] There are three important takeaways. First, students in Central Time have more sunlight before school than those in Eastern Time, as expected.

[11] A version of this graph with unconditional means for each group-time bin shows similar patterns.

Second, the cross-time zone movers neatly switch places as they move across the time zone boundary: the cross-time zone movers are now 'treated' with the start time of the other time zone. This shift allows us to identify the effects of start time relative to sunrise using only within-student variation. Third, the lines generally overlap within time zones, indicating that those who switch time zones are likely not selecting into schools in a way that affects sunlight before school.

More formally, Panel A of Table 4.2 presents the first stage regression of relative start times on time zone.[12] The first row displays the main effect for all students, and the second row displays the interaction effect for pubescent students. The third row is the p-value from a test for the combined significance of the effect for pubescents. Each specification includes individual and age-gender fixed effects. Column 1 has no additional controls. Column 2 adds longitude.[13] Columns 3 and 4 add demographic means at the district and school level, respectively. These demographic means include the percentage of students who are male, FRL, black, Hispanic, and Asian. Columns 5 through 7 are identical to Columns 2 through 4, but with the addition of indicator variables for 1, 2, and 3+ years after the move to account for potential disruption.[14]

All specifications yield similar estimates. We prefer Column 7 because it includes controls that address both disruption and potential changes in peer characteristics over the move. Across the columns, younger children in ET have about 25 fewer minutes of sunlight before school than

[12]The Online Appendix includes robustness checks using additional controls including urbanicity, log income, school size, student/teacher ratio, and other levels of demographic aggregation. The results are similar to Table 4.2.

[13]We also consider adding latitude as a control. However, our study area has a relatively small north-south dimension — from the top to the bottom of the panhandle, the difference in average sunrise time over the school year is less than a minute. When we include latitude as a control, the main results are very similar but slightly smaller in magnitude. These robustness checks can be found in the Online Appendix.

[14]We consider specifications that control for the time until the move. This has almost no effect on the other coefficients in both the first and second stage, but we do not pursue this avenue to avoid controlling for information that the students may not have themselves.

children in CT, while those who have gone through puberty have about a 41 minute difference. The difference is less than 60 minutes for each age group, which is what we would expect if schools opened at the same clock time on either side of the time zone boundary. We take this as evidence that policymakers faced with later sunrise times may shift start times later to compensate, and that they may differentially shift elementary start times to prevent younger students from waiting for the bus in the dark.[15] The F-statistics for the first stage range from 825 to 2004, with an F-statistic of 1105 for our preferred model.

1.5.2. Effect of start times on academic achievement

Panels B and C of Table 4.2 contain estimates for the effect of relative start times on math and reading test scores. Each specification includes individual fixed effects and age-gender dummies,[16] and the columns add additional controls in the same order as Panel A.

In Panel B, the estimated effect of relative start times on math scores is similar after we add a control for longitude in Column 2. In all subsequent specifications, moving start times one hour later increases math scores for prepubescents by 0.009-0.020 SDs; none of the coefficients are close to statistically significant. For adolescents, later start times increase math scores by 0.077-0.084 SDs. Across specifications, both the adolescent level and the difference between adolescent and prepubescent scores is significantly different from zero at the 1% level.[17]

[15] When we look at results by age, the difference in sunlight before school is 22-23 minutes for elementary school students (typically ages 8-10 in our data), 28-30 minutes for middle school students (ages 11-13), and 47-59 minutes for high school students (ages 14-15).

[16] Test scores are normalized at the year-grade level, so if we included the entire state population the age-gender dummies would reflect only the age-varying gender gap. Because our sample is restricted to movers in the Florida panhandle, there may be additional age-varying differences relative to non-panhandle and non-mover students that the age-gender fixed effects pick up. They are particularly important to include because they function as a set of saturated dummy variables for puberty, which we interact with start time as a explanatory variable of interest.

[17] The difference in effect size by pubertal stage is striking, and corresponds with increasing sensitivity to start times during puberty. In Online Appendix Table A6 we estimate a version of Table 4.2 without the interaction.

Panel C repeats the exercise for reading. The results are again consistent across the columns; in our preferred specification moving start times one hour later increases reading scores by 0.061 SDs for prepubescent students and by 0.057 SDs for adolescents. The overall effect for adolescents is statistically significant at the 1% or 5% level for all specifications, while for prepubescents it is either significant at the 5% or 10% level depending on the level of aggregation for the demographic controls. There is no statistical difference between pubertal and prepubertal effects. For adolescents, the effect size is larger in math than in reading across specifications, corroborating previous research on middle schoolers (Edwards, 2012; Ng et al., 2009).

1.5.3. Mechanisms

There are (at least) two reasons why school start times might affect academic achievement. First, later start times relative to sunrise may make it easier to get to school on time, reducing absences and increasing time spent on instruction. Alternatively, more sunlight before school may improve cognitive function by increasing sleep levels and alertness.

Panel D of Table 4.2 explores the relationship between start times and absences. Conditional on school or district level demographic controls, there is no statistically significant relationship between start times and absence rates for adolescents, although there is an estimated 0.9% decrease in absences for the younger students in the preferred specification. For all ages, later relative start times decrease absences, although the relationship is weaker for adolescents than for prepubescents, which is difficult to reconcile with the larger effects of start times on achievement we observe in math and reading. Comparing between age groups is somewhat fraught; because record-keeping is not standardized across schools, an elementary-aged child might be

---

The average effect of start times on achievement is close to the average of the adolescent and pre-pubertal measures; the reading estimates are statistically significant but the math estimates are only sometimes statistically significant.

marked absent for the entire day when she is late in the morning, but a high schooler who is similarly late could be marked absent only for the first class but not as absent in the larger tracking system. However, that caveat addresses only differences between the age groups; in light of the moderate and imprecisely estimated effects on absences for all age groups we think it is unlikely that reductions in absences are a major causal channel through which later relative start times translate into improved test scores.[18]

The evidence is somewhat stronger in favor of sleep and alertness as the causal channel. Our data do not contain information on sleep, so we use the Child Development Supplement (CDS) of the Panel Study of Income Dynamics (PSID) to estimate the effect of the time zone boundary on sleep. The CDS collected time use diaries for students in 1997, 2002, and 2007, along with geographic and demographic information. We regress hours of sleep on a dummy variable for residence in ET for children within 400 miles of the CT-ET boundary.[19]

Table A7 in the Online Appendix shows that prepubescent children in ET get 6 minutes less sleep per night during the week than children in CT.[20] The difference in sleep is reversed on the weekend as they attempt to correct the sleep deficit; students in ET sleep 4 minutes *more*. After the onset of puberty, both gaps widen: children in ET get 17 minutes less sleep per night during the week, and compensate with 13 minutes more sleep per night on the weekend.

These findings indicate that children in ET are more sleep-deprived than children in CT, and that this gap increases in adolescence. If school start times in our Florida sample are

[18] We do not have data on tardiness, which could also be affected by start times.

[19] The publicly-available CDS does not geocode individuals at a sub-state level, so we exclude all observations from states with multiple time zones — including Florida. See the Online Appendix for more information on sample construction.

[20] All estimates reported here include demographic controls; see Column 2. We conservatively cluster by state. The difference in sleep between children in ET and CT is statistically different for adolescents but not for prepubescent children.

representative of start times elsewhere, this suggests a passthrough from relative school start times to sleep of 40-50%, which is comparable to the 46% found by Wahlstrom et al. (1998). Thus, moving from ET to CT increases both sleep and test scores (and increases them more for adolescents), suggesting that levels of sleep and alertness in the morning are important causal channels through which later school start times increase achievement. There may be other changes in time use — descriptive research indicates that later start times decrease time spent on extracurricular activities, as well as reduce leisure time for girls and computer use for boys (Groen and Pabilonia, 2015; Wahlstrom et al., 1998) — but it is difficult to reconcile the patterns of achievement by developmental status with an explanation *not* revolving around the transition to puberty. More importantly, from the perspective of a policymaker the distinction is moot: whether the causal channel is before-school time or after-school time, changing the school start time will affect both channels.

1.5.4. Heterogeneity by age and gender

Rather than allowing the effect of relative start times to vary by pubertal status as in Equation 1.1, it is possible to estimate each age-gender-start time interaction term separately. If the increasing importance of start times for math performance is a function of puberty, the effect sizes should grow in importance as a larger share of the gender enters puberty. This is precisely what we see.

Figure 4.4 presents coefficients from a version of Equation 1.1 estimated separately by gender, with start time fully interacted with age. Because ages range from 8 to 15, this amounts to estimating

$$y_{it} = \sum_{a=8}^{15} \delta_a h_{it} \times \mathbb{1}[\text{age}{=}a] + X_{it}\beta + \gamma_i + \varepsilon_{it} \tag{1.4}$$

where $h_{it}\mathbb{1}[\text{age}{=}a]$ is instrumented by time zone interacted with age, and X_{it} is the baseline vector of controls. Starting in the upper left corner of Figure 4.4, there is a sharp spike in the effect of school start times on math scores at age 11 for girls, precisely when the median girl enters Tanner Stage 3. The effect of later school start times is statistically significantly different from zero for girls 11-13, but not for girls 10 or younger. For boys, in the upper right corner, the effect of start times on math scores is statistically indistinguishable from zero at the 10% level for ages 8 to 12, then jumps from 0.049 to 0.096 at 13 as the median boy enters Tanner Stage 3. The effect of start times is significantly different from zero at the 1% level for ages 14 and 15. This is evidence that the increasing importance of start times with age is driven by pubertal entrance, rather than other academic or behavioral changes.

The effect of start times on math scores is noticeably (though insignificantly) smaller for girls after age 13. One possible explanation is that certain stages of puberty are particularly important for sleep (Campbell et al., 2012), and girls have moved beyond this developmental stage by age 14. For example, Crowley et al. (2007) speculate that older adolescents may be less responsive to light than younger adolescents. However, there is no firm physiological evidence on sleep patterns or light sensitivity at a granular gender-age level, so resolution of this issue will have to wait for data which extends further into adolescence, especially for boys. There is persuasive evidence from Carrell et al. (2011) that start times have a large effect on achievement for college freshmen cohorts that include both boys and girls, so we think it is unlikely that the true effect is zero for 14 and 15 year old girls.

In reading, as one might expect from Table 4.2, there is no sharp change in the relationship between start time and achievement at the gender-specific puberty thresholds.

1.5.5. Heterogeneity by subgroup

Educational interventions often have a larger effect on disadvantaged students or students attending low-resource schools (see, e.g. Krueger and Whitmore (2002)). In this case, however, there are more similarities than differences in effect sizes across racial, economic, and gender groups. The standard errors are large, but the results suggest that changes to start times will benefit all students, rather than certain demographic groups.

In Table 4.3, we apply our baseline regression of test scores and absence rates on start times for each of six demographic subgroups: whites and minorities;[21] FRL and non-FRL; and male and female. In math, the effect sizes are similar between white and minority students in Columns 1 and 2. For pubescents, a one-hour delay in relative start times increases math scores by 0.093 SDs for white students and 0.081 SDs for minority students. In reading, the effect sizes are 0.040 and 0.132, respectively, though this difference is not significant. None of the estimated effects for absences are statistically significant.

Columns 3 and 4 contrast FRL and non-FRL students. The effect size for math scores is significantly larger for the non-FRL adolescents at 0.147 SDs per hour, compared to 0.048 SD for FRL adolescents. There are no statistically significant differences for reading scores or absence rates.

[21] We count all non-white students as minorities. These results are not substantively affected by not counting Asians as minorities, or delineating the categories as black and non-black. In the latter case, however, the standard errors for the black sample are large.

Finally, Columns 5 and 6 indicate that the effect of relative start times on achievement and absences is similar for boys and girls. The difference is never statistically significant, and the effect sizes for both groups are similar to the overall estimates of Table 4.2.

1.5.6. Persistence of start times

To this point, we have not distinguished between a transitory and permanent effect of start times on academic achievement. This distinction could be important. If changing school start times from one year to another has an effect for (say) only one year while the student adjusts her sleep schedule, our estimates (which are essentially the average of achievement before and after the move) would overstate the long-term effect by averaging a positive effect in the first year with a zero effect in all other years. This would mean that our estimates would not correctly predict the long-term change in achievement as a result of changes in start time policy. We explore this possibility in Table 4.4, where we estimate a version of our baseline regression with relative start time by pubertal status interacted with dummies for 1, 2 and 3+ years since move. Note that the pubescent effect is the total estimate for adolescents, rather than an interaction.

The results indicate that the short-term and long-term effects are quite similar; for pre-pubescent children the long-term math and reading coefficient is an insignificant 0.005-.011 SDs smaller. For adolescents, the math effect is 0.020 SDs lower in the long run; the difference is significant at the 1% level. The reading effect is 0.010 SDs higher in the long run; the difference between the short and long run is not statistically significant. In the long run, the effect is larger for adolescents than younger students in both subjects, although the difference is not statistically significant in reading. In both the short and long run, the adolescent effects differ from zero. We conclude that changes to start times improve math and reading achievement

within a year of the change in sunlight exposure for adolescents, and the effects largely persist over time.

1.5.7. Learning versus testing

The positive effect of later relative start times on test scores has two potential causes: improved learning in the year leading up to the test, or better testing performance caused by increased alertness on the day of the test. Our approach so far has been to estimate the combined effect of learning and testing. Fully disentangling the two effects would require separate instruments for start times during the year and on the day of the test, which are unavailable in our data.

The data allow us to answer a related but less definitive question: does the relationship between sunlight and achievement vary with the amount of baseline test-day sunlight, holding sunlight during the school year constant? If so, this implies that changes to test-day relative start times matter for achievement. Estimates of the marginal effect of later relative start times at different levels of test-day sunlight can be combined with a mild assumption of diminishing returns to sleep to generate a lower bound on the size of the test-day start time effect.

This strategy is possible in our context because our data contain variation in test-day relative start time that is separate from the cross-time zone variation in start times. During the study period, testing dates moved from late February to mid-April. This changed levels of sunlight on the day of the test, but had only a small effect on average sunlight levels during the school year when learning occurred. Using these policy changes, we find that the lower bound on the test-day effect is relatively high for reading, but low for math. We interpret this as evidence in favor of potential testing effects in reading, but not as a definitive rejection of testing effects in math.

During the study period, the FDOE pushed the testing period later in two discrete steps. The first change was particularly useful for this research, because it moved the testing period from before to after the start of Daylight Saving Time. DST begins with a time change on the second Sunday of March in most of the United States.[22] Clocks "spring forward," moving sunrise one hour later and reducing the amount of sunlight before school. Figure 4.6 charts sunlight before school for 2000-2007, 2008-2009, and 2011-2013, corresponding to the three test-day policy eras.[23] In 2000-2007, testing took place just before the change to DST, meaning that there was a relatively large amount of sunlight before school; in ET, the average was 1 hour 20 minutes on the first day of testing. For 2008 and 2009, the test was moved two weeks later to directly after DST; the average amount of sunlight before school on the morning of the test in ET dropped to 28 minutes. In 2011, the test was moved one month later, increasing sunlight before school on the testing day to an average of 1 hour 9 minutes for 2011-2013.[24] Throughout the study period, the average sunlight before school in the school year leading up to the test barely changed, at 61, 56, and 59 minutes, respectively. Based on these differences, we group together 2000-2007 and 2011-2013 into a "late test time" treatment, and 2008-2009 into an "early test time" treatment.[25] As the testing date was moved back, preparation time increased for all students; however, because the early test time treatment occurred in the middle of the period (when the testing date was closest to the DST transition) the average preparation

[22]There have been changes in DST dates in the recent past; before 2007 DST started on the first Sunday of April. This change is not relevant for this research, because testing occurred before DST began in all years before the switch in DST dates.

[23]Specifically, the Figure shows 2007, 2008, and 2011, but all are archetypes of their eras.

[24]We exclude 2010 from analysis in this section because DST occurred during the testing period in this year, meaning that we cannot assign the test to either pre- or post-DST. We also exclude 1999 because testing occurred one month earlier, in the first week of February, where the sunrise time is between the early and late period.

[25]The main difference between 2000-2007 and 2011-2013 is that the average relative start time in the year preceding the test was slightly earlier in 2011-2013 because the extra month of class time was almost entirely after the DST transition. Excluding 2011-2013 from the regressions does not change our conclusions.

time is only five days longer for the late test time treatment group. Furthermore, neither of the changes in testing date correspond to any major changes in testing procedure or curriculum we could find, suggesting that any differences in performance between the policy eras can be attributed to test-day sunlight.

It is tempting to estimate the effect of earlier relative start times on the day of the test by regressing test scores on a dummy variable for the testing era. However, test scores are standardized by the mean statewide score in each grade-year, so direct comparisons between years are not possible. We instead test whether the effect of full-year relative start times on achievement changes depending on test-day sunlight. We estimate a second stage of:

$$\begin{aligned} y_{it} = {} & \phi_1 hours_{it} \mathbb{1}[\text{child} \cap \text{late test time}]_{it} + \phi_2 hours_{it} \mathbb{1}[\text{child} \cap \text{early test time}]_{it} + \\ & \lambda_1 hours_{it} \mathbb{1}[\text{puberty} \cap \text{late test time}]_{it} + \lambda_2 hours_{it} \mathbb{1}[\text{puberty} \cap \text{early test time}]_{it} + \\ & X_{it}\beta + \gamma_i + \varepsilon_{it} \end{aligned} \tag{1.5}$$

where X_{it} includes, in addition to the usual controls, dummies for the policy eras and their interaction with puberty.

Because sunlight before school during the year leading up to the test is nearly identical between eras, the difference in coefficients for a given age group represents the change in the effect of one extra hour of test-day morning sunlight on test scores between two different margins: 1 hour 17 minutes from sunrise (the average in the late testing years) and 28 minutes from sunrise (the average in the early testing years). If the coefficients are the same, that implies either that the effect of test-day sunlight is identical at the two margins, or that the effect of test-day

sunlight is zero.[26] If they are different, that implies there is some effect of testing day sunlight on at least one of the margins. A smaller coefficient in the late-testing years is consistent with diminishing marginal returns to test-day sunlight.[27] Analogously to the main specification of Equation 1.1, we expect that $\lambda > \phi > 0$, since later start times should improve performance more for adolescents than for younger students.

Table 4.5 presents our findings for math and reading. Unlike the main table, the coefficients estimate the full effect for adolescents, rather than the difference between adolescents and younger children. We begin by verifying in Columns 1 and 3 that excluding 1999 and 2010 does not substantively affect our baseline results.

Columns 2 and 4 estimate Equation 1.5, allowing for a differential effect of start times on achievement as a function of baseline test-day start times. In Column 2, the math results are unchanged from our main specification: moving relative start times one hour later increases achievement at a similar rate in the two eras for adolescents (0.096 SDs per hour in the early versus 0.095 SDs in the late era), and the difference in estimates is statistically insignificant. Because we argue there should be diminishing marginal returns to more sunlight before school, we take the similarity in estimates between different test-day sunlight eras as evidence against test-day effects in math.[28] For children, the results are slightly more suggestive of testing effects,

[26]The latter implication is technically a subset of the former, but the conceptual difference is important.

[27]Formally, this can be seen by modeling outcomes y as an additive function of full-year and test-day sunlight, $y = f_{year}(t_1) + f_{test}(t_2)$. We estimate $\beta_{early} = f'_{year}(56m) + f'_{test}(1h\ 17m)$ and $\beta_{late} = f'_{year}(56m) + f'_{test}(28m)$, where f' is the first derivative. Then, $\beta_{early} - \beta_{late} \approx f'_{test}(1h\ 17m) - f'_{test}(28m)$, so a positive difference is evidence for diminishing marginal returns. Any non-zero difference implies that the function relating performance and testing has a non-zero effect at (at least) one of the margins.

[28]Technically, the similarity between the early- and late-test time coefficients cannot be read as a failure to reject testing as an important input into math achievement. It is instead a rejection of a nonlinear relationship between achievement and test-day sunlight — it is consistent with an effect of test-day start times only if the relationship between achievement and start times is linear in the region between 28 minutes and 1 hour 17 minutes of sunlight before school.

with larger effects for more sunlight on the test days with less sunlight before school (at 0.071 SDs per hour) than on the test days with more sunlight before school (at 0.022 SDs per hour). However, neither estimate statistically differs from zero, nor do they differ from each other.

In reading, the results are more strongly suggestive of testing effects. For younger children, one extra hour of morning sunlight increases test scores by 0.096 SDs in years with less sunlight before school (early years), while the effect is statistically insignificant and only 0.049 SDs in years with more sunlight before school (late years). For adolescents, the effect during the relatively earlier testing era is 0.104 SDs per hour of sunlight, compared to 0.045 SDs in the late era. The difference in estimates is statistically significant for adolescents, suggesting that test-day sunlight may be important for reading achievement. Under the assumption that changes to test-day relative start times do not change the effect of start times during the school year, and that there are decreasing marginal returns to later test-day start times, this indicates that the test-day effect is bounded at a minimum of 0.059 SDs per hour for adolescents (calculated as 0.104-0.045) and 0.047 for prepubescents (0.096-0.049) in the early start time years. This bounded effect implies that testing is a more important causal channel than learning for reading achievement.

There is, however, one important reason why the result in reading should be taken with some caution. In both of the early-testing years, the testing period began almost immediately after the switch to DST; one day after in 2008 and two days after in 2009. Because clocks move forward during the spring DST transition, students can lose up to an hour of sleep, depending on how much they adjust their sleep times. There is strong evidence that the DST transition negatively affects sleep levels and alertness: Smith (2016) finds an increase in the number of fatal car accidents in the six days following DST. We therefore interpret the difference in

coefficients between the early- and late-baseline years as the difference in the gains from an hour of sunlight on test-day with a baseline of 1 hour 17 minutes sunlight before school and the gains from an hour of sunlight on test-day with a baseline of 28 minutes of sunlight before school *and* up to an hour of sleep deprivation. We have no information on the testing date for each student, so we cannot further stratify the start time effect as a function of number of days since the DST transition. However, since the testing period was longer than one week in both 2008 and 2009, the test was likely taken a few days after the DST transition and perhaps as long as two weeks after, when transition-induced sleep loss has lessened. We therefore think that the safest interpretation is for moderate test-day effects in reading, of the same order as the full-year learning effects. At the very least, this result tells us that under an assumption of diminishing marginal returns to test-day sunlight, there are some situations (potentially including more sleep deprivation than is normal for this age group) where test-day sunlight has a large effect on academic achievement in reading. There is much more to be done to separately identify the effects of whole-year and test-day sunlight, but we leave this for future research.

1.5.8. Placebo time zone changes

The identification strategy in this paper leverages the discontinuity in sunrise times at the time zone boundary to estimate the effect of relative start times on academic performance. In a reduced form sense, we track students as they move east (west) over the time zone boundary in the Florida panhandle and find that scores decline (increase), as predicted by the earlier (later) relative start times.

Alternatively, perhaps moves to the east are score-decreasing for some reason unrelated to start times: schools are lower quality, or parents moving east get worse jobs and lower pay,

which decrease investment in educational inputs. Our baseline specification includes controls for longitude and school demographics, which together control for any variation in underlying school or family characteristics that is linearly correlated with the demographic controls or varies linearly from east to west over the panhandle. If there are nonlinearities in this relationship, however, our method could misattribute variation in unobserved non-start time inputs to variation in start times, biasing our estimates.

In this section, we estimate placebo regressions that attempt to rule out a non-start time explanation. We generate placebo boundaries in ten mile increments from the true boundary; Figure 4.8 displays the estimated effect of moving over each placebo boundary, conditioning on true time zone, the regular vector of controls, and student fixed effects. We present estimates using cross-time zone movers, as well as restricting to only within-time zone movers. In Section A.1.8 of the Online Appendix, we demonstrate that schools very close to the time zone boundary adopt start times similar to their cross-boundary counterparts; this means that there is a treatment effect of moving to or from the region directly adjacent to the boundary, even when the move is within time zone. We therefore exclude a 25 mile area around the true boundary (a version of the placebo test without this exclusion is available in the Online Appendix).[29]

Figure 4.8 displays the estimated coefficients for moving over placebo boundaries, placed in 10 mile increments from the true time zone boundary. In math, the placebo coefficients for the adolescent interaction are always smaller than the true coefficient, and usually significantly so. The true level coefficient is approximately zero, and the placebo coefficients bounce around

[29]Excluding this region is not necessary in the main specification, as the IV estimate accounts for treatment bleed across time zones. However, our results are substantively the same even excluding this donut; we estimate that moving start times one hour later would improve math scores by 0.065 SDs for adolescents, and would have little effect on prepubescent math scores or reading scores for either age group. The number of students also decreases, resulting in larger standard errors on these estimates.

that estimate, although we can sometimes reject they are zero. In reading, for both the within- and all-mover specifications, the placebo coefficients are almost always smaller than the true coefficients (and very imprecisely estimated when they are not). The true time zone-puberty interaction coefficient is approximately the same size as the placebos, although it is imprecise enough that we cannot differentiate it from zero in our main sample.

In summary, we estimate regressions of outcomes on placebo time zones, and find little evidence of changes in outcomes over the placebo boundaries, suggesting that the gains in achievement from westward moves are a function of crossing over the true time zone boundary and being exposed to later relative start times, rather than improvements in some other input.

1.5.9. Other effects of cross-time zone moves

A final threat to our identification strategy is the possibility that moving between time zones has a direct effect on family income or other characteristics. If these changes have an independent effect on academic performance, the exclusion restriction would be violated. Gibson and Shrader (2015) show that a one-hour delay in sunrise time reduces wages by between 0.5 and 4.5%. Given Dahl and Lochner's (2012) estimate of a 0.06 SD decrease in test scores per $1,000 decrease in EITC income, this could explain much of the test score effect. We do not observe parental income, and so cannot directly control for this possibility. However, there are three reasons to expect that a measure of income is not an important missing variable in our analysis. First, jobs are a primary reason for moving long distances and are chosen by the parents; wages are an important factor in job choice. It is therefore unlikely that movers are immediately treated with the average difference in wages given the change in sunrise times over the move. In fact, Gibson and Shrader (2015) argue that housing prices adjust to eliminate the incentive to move,

and document that housing is indeed more expensive in early-sunrise cities. Disposable income would then be flat over the move, eliminating any effect on academic achievement. Second, in our sample zip-level income is higher in low-sunlight ET than in high-sunlight CT, which is the opposite of what is predicted by Gibson and Shrader (2015).[30] As we demonstrate in the Online Appendix, our results are unchanged by controls for zip-level income. Third, and most importantly, even if disposable income did increase as families moved over the time zone boundary, we would expect that children of all ages would benefit from the move. Instead, we observe larger increases in standardized test scores for pubertal children — and almost no increase for pre-pubertal children in math — suggesting that changes in sunlight before school are the most important causal factor.

1.6. Benefits of rearranging start times

Academic research and popular coverage of the potentially negative effects of early start times dates back at least as far as the late 1990s (Douglas, 1999; Wahlstrom et al., 1998). The evidence from the medical and physiological literature has grown so compelling that the American Academy of Pediatrics now recommends that middle and high schools delay start times to allow students to wake up no earlier than 8:00 am (2014). Despite the growing consensus, schools continue to open early; the median high school *opens* at 8:00 a.m. (NCES, 2012).

School districts, particularly those in large urban areas, often open different types of schools at different times. This structure is convenient for parents dropping off children at different schools, because it guarantees that a child in middle school will not need to be dropped off at the same time as a child in high school. It also allows school districts to use the same

[30]This does not seem to be a function of education, since literacy is actually marginally lower in ET (Authors' calculations from the NCES 2003 National Assessment of Adult Literacy).

buses more intensively, saving on transportation costs. However, of the 19 school districts in the Florida panhandle, only 4 currently order their start times in the 'efficient' way. Inflexible parental schedules often preclude moving start times later for all students, since parents must be able to drop off their last child in time to get to work. In this section, we consider the academic effects of an alternative start time policy that better fits the physiological evidence but does not alter the overall distribution of start times: changing the opening order for different types of schools to elementary schools, middle schools, and finally high schools.

We operationalize this simple counterfactual by taking the average start time for each school type in each district, then assigning the earliest average start time to elementary schools, the next start time to middle schools, and the latest time to high schools. We adjust the mean start time for each district so that it is the same in the counterfactual as in the real world.[31] We take the difference in relative start times for the counterfactual and real worlds for each school type and apply the coefficients from Table 4.3, weighting by the number of children in each district-school type. On average, this moves elementary start times 22 minutes earlier, middle schools 13 minutes earlier, and high schools 44 minutes later.

Figure 4.10 displays the effect on test scores, separated by gender and race. The counterfactual policy has been constructed so that if start times have an identical effect on children of all ages, the average increase in test scores will be zero. However, because the gains from later start times are smaller for younger children than for older children, our procedure has the effect of raising average academic achievement. In both math and reading, the effect is slightly (and

[31] A clarifying example: if a district has 800 students in grade 9-12 schools with a start time of 7:00, 800 students in grades 6-8 schools with a start time of 7:30, and 1200 students in K-5 schools with a start time of 8:00, the mean district start time is 7:34. We would then set counterfactual start times to 7:08 in elementary school, 7:38 in middle school, and 8:08 in high school, with an average start time of 7:34. The procedure keeps the counterfactual mean start time the same as the status quo, and maintains the half hour spread in start times between school types.

usually insignificantly) negative for all groups of students in elementary and middle school. On average, elementary- and middle-school math and reading scores decline by 0.01 SDs. For high school students, the gains are large and statistically significant: in math, the proposed policy would increase minority student achievement in high school by 0.06 SDs in math and 0.08 SDs in reading. For white students, we expect that math scores would increase by 0.06 SDs and reading scores by 0.02 SDs. By gender, male high school students benefit slightly but insignificantly more compared to females. Using the coefficients from Table 4.2, the average effect is a 0.064 SD gain in math and a 0.044 SD gain in reading.

Furthermore, the high school results are good estimates for the overall change in achievement for each student by the end of high school. In Section 1.5.6 we show that increases in academic achievement occur immediately after the move and persist for years. That implies that back-loading the later start times will increase achievement as of the conclusion of high school by approximately the same amount as the single-year effect. Alternatively, taking the long-term estimates of Table 4.4 as given, the counterfactual would increase end-of-high school math scores by 0.05 SDs and reading scores by 0.037 SDs.[32]

One drawback of re-ordering start times would be that the youngest children may have to wait for the bus or walk to school in the dark. In December, the average sunrise would be only 53 minutes before school starts, with 12% of elementary school students having less than half an hour between sunrise and school start in the darkest month. This would likely mean that a substantial number of very young students might need to travel to school in the dark, which presents a significant drawback to this proposal. Moving all school start times later, rather than re-ordering schools, would not have this problem.

[32]This math score is calculated by multiplying the long-term coefficent of $0.087 - 0.020 = 0.067$ by the average change in high school relative start times, 44 minutes.

In summary, we demonstrate that adjusting school start times so that high school students have the latest start time would significantly increase achievement for older children at a very low academic cost for younger children. Even when start times are reordered such that the average start time across the district remains the same, there are non-trivial gains in average academic performance that would benefit students in all demographic groups. These gains must be weighed against the costs of having younger children traveling to school in the dark.

1.7. Conclusion

We investigate the effect of daily school start times on academic performance. Adolescents in particular struggle with early start times; the onset of puberty shifts the sleep schedule back several hours, making any given start time more onerous for high schoolers than for students in other age groups. Our empirical strategy tracks academic performance in the same student before and after a cross-time zone move, which we use as an instrument for the amount of sunlight before school. Because the circadian rhythm is tied to variation in sunlight levels, this is a good approximation of a policy change in start times. Using a long individual panel from the state of Florida, we find that moving start times one hour later relative to sunrise would increase adolescent scores by 0.081 SDs in math and 0.057 SDs in reading. The increase in test scores can be observed immediately after the move, and persists for as long as we can measure it. Taking advantage of the fact that girls enter puberty two years earlier than boys, we document that the effect of relative start times on math performance spikes precisely at the gender-specific age of median entrance into an important pubertal stage. Previous research, which has mostly focused on a smaller age range of the population, has been unable to fully explore changes in the effect of start times over the pubertal transition.

These effects are cost-effective compared to other proposals to improve educational achievement, such as smaller classrooms or higher-skilled teachers. Specifically, reducing class size in elementary schools from 22 to 15 increases scores by 0.15-0.20 standard deviations (Schanzenbach, 2006), and a 1 standard deviation improvement in teacher quality increase scores by approximately 0.10 standard deviations (Chetty et al., 2014). Changes to school schedules would likely be much cheaper. Jacob and Rockoff (2011) suggest that the cost of moving start times one hour later is less than $150 per student per year and potentially as low as free. In contrast, reducing class sizes by a third costs approximately $6,200 per student per year.[33] The cost of such a large improvement in teacher quality is more difficult to evaluate, since the supply side of the teacher market is poorly understood. However, it is likely very large, if only because it would likely require hiring hundreds of thousands of new teachers.[34]

We simulate the effect of adjusting start times by school type to match students' developmental patterns while maintaining the same mean district start time. We estimate that this would increase math scores for high school students by 0.064 SDs and reading scores by 0.044 SDs, while having small and mostly statistically insignificant effects on scores for younger children. Alternatively, moving start times later across the board would increase achievement for all ages and demographics. In either case, adjustments on the start times margin seem to be significantly cheaper than adjustments to classroom size or teacher composition, suggesting that there may be large unrealized gains in this area.

[33]These figures are from Schanzenbach (2006), inflated from 2002 to 2011 prices via the CPI.

[34]If teacher quality were distributed normally, then replacing the bottom half of teachers with average teachers would raise the average SD of teacher quality by only 0.4, and therefore test scores by 0.04 SD. According to the NCES, there were 3.7 million teachers in the United States in 2012. It is hard to imagine that finding 1.85 million new average-quality teachers could be done without significantly increasing wages.

There is one important caveat to our findings. Changes in school start times can increase achievement through either better learning in the year leading up to the test, or improved testing performance. We exploit a policy change in the testing date relative to Daylight Saving Time to learn whether test-day start times are important for achievement (but not by how much). We find suggestive evidence in favor of testing effects in reading, but not math. Our method is unable to precisely quantify the relative importance of testing and learning, but show that the magnitude is approximately the same for reading. We leave this as an important direction for future work.

Despite growing medical and physiological evidence that current school start times are too early for optimal adolescent cognitive functioning, there has been little policy response to move start times later. We add to this debate with direct evidence that more sunlight before school — or a later relative start time — increases academic achievement for children of all ages. The increase in scores is much larger for adolescents, implying that even when parental schedules preclude later start times for all children, districts can improve academic performance by adjusting the order in which school types open to correspond with students' changing sleep schedules. Specifically, high school students should begin school later in the day to compensate for pubertal changes that shift their circadian rhythm later, while elementary students should begin school the earliest. Despite the low costs of adopting this policy, the gains are quite large.

CHAPTER 2

Judicial Errors: Evidence from Refugee Appeals

2.1. Introduction

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

The efficiency and fairness of the courts has far-reaching consequences. Beyond the literally life-altering effects of decisions on criminal defendants, Porta et al. (1998) argue that inefficient civil courts increase transaction costs and reduce aggregate investment. However, evidence on the overall fairness of the courts remains limited. In specific situations, there is compelling evidence that judicial decisions are affected by non-relevant factors like upcoming elections (Canes-Wrone et al., 2014), inter-communal violence (Shayo and Zussman, 2010), television news about unrelated crime (Philippe and Ouss, 2016), the previous decision (Chen et al., 2016), the timing of the hearing relative to lunch (Danziger et al., 2011) and the winner of last night's football game (Eren and Mocan, 2016). All of these findings suggest that some cases would receive a different outcome if their case had been heard on a different day or even at a different time. However, aggregating all these different ways a court can be unfair (and others yet to be

uncovered) is unlikely to be feasible. In this paper, I take a different approach and introduce a new method to measure overall unfairness in judicial decision-making, as well as the reliability of individual judges.

I begin with a simple definition: a court is *fair* if decisions are made according to the facts of the case and the universal application of the law — or equivalently, if all decisions are made according to the same standard. Judges can differ from this ideal in two ways. First, judges could differ in terms of aggregate leniency, and in fact in all courts that I am aware of there is substantial cross-judge variation in propensity to incarcerate randomly-assigned defendants (Abrams et al., 2012; Aizer and Doyle, 2013; Bhuller et al., 2016; Mueller-Smith, 2014). Second, judges with the same overall incarceration rate could choose to incarcerate different individuals, and judges might make different decisions today than they would yesterday. This latter form of unfairness, which I refer to as *inconsistency*, is implied by the examples from the previous paragraph. The tools I develop in this paper allow me to separately identify judge-specific leniency from a measure of overall consistency that encompasses all the above examples, as well as any other decisions where a judge violates his own standards. Together, these estimates facilitate evaluations of the efficiency of the judicial system (are guilty defendants likely to be incarcerated?), and allow the researcher to measure the quality of individual judges and the success of reforms. I show that the presence of inconsistency implies violations of the montonicity assumption in examiner-assignment IV designs, and my results shed light on potential biases in this increasingly-common identification strategy (Dahl et al., 2013; Mueller-Smith, 2014).

My conception of consistency is a simple generalization of the usual index model of judicial decision-making, where judges perfectly observe the strength of each claimants case and

approve them if it is larger than some judge-specific threshold.[1] In my model, judges observe case quality with error; the size of the distribution of this error is inconsistency. Judicial behavior is thus summarized by a judge-specific threshold and a judge-specific error distribution. In many environments this is not identifiable: if we observe only judge-specific approval rates, a judge who perfectly selected all claimants meeting the legal standard would be indistinguishable from a judge who approved the same share of claimants but flipped coins to do so.

Identification relies on two distinct institutional characteristics. The first is random assignment of judges to cases, which is common in many court systems and ensures that underlying case characteristics are uncorrelated with judge characteristics. The second is that the decision is made by two judges acting independently but using similar criteria. This latter condition can be met by identical standards for each judge, but in my setting and in many others is satisfied by the requirement that claimants be recognized as having an arguable case by one judge (in legal parlance, granted leave) before being given a full hearing in front of another. I use the second-stage decision to check the accuracy of the first-stage decision — if there are two first-round judges who approve the same number of first-round claimants, the more consistent judge will have a higher share of her claimants approved by the second-round judge.[2] In my context, consistency maps directly into judge quality: judges are specifically told that they should approve first-round claimants who can make a strong case in the second case.

A similar comparison allows identification of second-round consistency: approval rates for consistent second-round judges — who can easily distinguish between high and low quality

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

[2]This also implies that the two first round judges would approve different individuals in the first round. I discuss the relationship between inconsistency in this sense and calculating the amount of disagreement between two judges in subsubsection 2.2.3.3.

cases — increase more than for inconsistent judges when the severity of the first-round judge (and corresponding case quality of the approved) increases. I show that the rest of the model, including the distribution of unobserved case strength, can also be identified using regressors that affect judge leniency but not errors. I build a structural model combining the two sources of identification that identifies leniency and consistency for each judge, as well as the distribution of underlying case quality. The model is nonparametrically identified, and can be tractably estimated via maximum likelihood under parametric restrictions.

This paper is related to two different literatures. First, the idea behind my identification strategy — that observing multiple decision-makers on the same case is informative about the accuracy of decisions — appears in other contexts. In a reduced-form sense, Frakes and Wasserman (2014) use patent decisions from non-US patent offices to generate an independent measure of patent quality, then examine how the quality of granted patents for US examiners changes as they are given less time to make a decision. Another set of papers grapples with selection-type models where outcomes are observed only conditional on treatment or some other agent decision. Analogously to how second-stage outcomes in my setting are observed only if the case is approved by the first-stage judge, Chandra and Staiger (2011) develop a model where hospitals both vary in their ability to treat heart attack patients, and choose which ones to treat. They identify hospital-level treatment effects using a structural model that interprets patient survival measures through a lens of treatment selection. Abaluck et al. (2016) study how doctors choose which patients to send for imaging tests for pulmonary embolism. Since the test reveals whether the patient actually has the disease, high test yield rates (conditional on share of patients sent for a test) are an indication of good allocation of tests across patients. Similarly, Anwar and Fang (2006) develop a hit rate test that compares the proportion of black and white drivers who are

found to be transporting drugs after a vehicle search to test for racial bias in the search decision among police officers. Closer to my context, Alesina and Ferrara (2011) use appeals in capital sentencing to test for racial bias under the assumption that higher courts are less racially biased than lower ones. I expand on this work by showing how the consistency of both the first- and second- round decision-maker can be identified, even when the researcher does not have access to objective measures of the truth. The model is applicable to any situation where two potentially fallible decision-makers are independently making a similar decision. It can be used to understand which decision-makers are most consistent, and what factors increase consistency.

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

problem is minor compared to the contribution made by judges who cannot be so characterized." In terms of my model, this means that there were high levels of inconsistency among these judges.

I apply my model to judicial review of refugee decisions at the Federal Court (FC) of Canada. The FC is the only point of appeal for claimants who have been rejected for refugee status by administrative decision-makers at the Immigration and Refugee Board (IRB), and is seen as a crucial backstop that ensures the fairness of the overall refugee system.[3] The stakes are high. As noted in Rehaag (2012), "if errors in first-instance refugee determinations at the [IRB] are not caught and corrected through judicial review, refugees may be deported to countries where they face persecution, torture or death." The judges are experts in dealing with refugee cases; about 70% of their caseload is refugee appeals. Nonetheless, I find low levels of consistency between judges, corresponding to a meaningful impact on decisions and outcomes. If judges were perfectly consistent, cross-judge variation in leniency would mean that judges would disagree on the correct decision for 8% of cases. With the addition of inconsistency, the disagreement rate jumps to 23%, implying that inconsistency is a larger contributor to unfairness than cross-judge variation in standards.[4] Furthermore, inconsistency seems to be caused by idiosyncratic observational errors, rather than permanent cross-judge differences in racial or gender bias, judging ideologies, or statutory interpretation.

The lack of consistency can also be understood as a failure of first-round judges to pick the claimants that are most likely to be successful at a full hearing in the second round. If all

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

[4]An alternative way of looking at this is that among judges with the same approval rate, they disagree on the correct decision 13% of the time.

claimants were given a full hearing, my results suggest that 19.4% of IRB denials of refugee status would be overturned, rather than the 6% that is ultimately successful under the current system.[5] This difference amounts to approximately 7,700 families over my study period.

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

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

In a final section, I show how my model can be used to construct counterfactual judge assignment regimes that minimize workload while approving the same number of claimants and maintaining the case quality of the approved. I find that the Federal Court could reduce refugee workload by approximately 18% while approving similarly-qualified claimants, saving at least $4.4 million in judge salaries alone over my study period.[6]

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

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

The paper proceeds in five parts. In Section 2.2 I present the model and discuss identification. Section 3.3 contains a discussion of the institutional background and data, and Section 3.5 the results. Section 3.6 concludes.

2.2. Model and identification

I discuss the institutional setting in detail in Section 3.3. To fix ideas before presenting the model, the Federal Court hears appeals from claimants who have been denied refugee status by the government. Enormous variation in approval rates for government decision-makers suggests that there is a long tail of claimants who would have been approved had they been assigned an alternative decision-maker, and should be successful on appeal. Decisions at the Federal Court are made in a two-stage process, where the criteria for a first-round decision is whether a claimant would have an arguable case in the second round. In this sense, the criteria are the similar in both rounds. The second round proceeds only if there is a first-round approval, and includes a hearing where lawyers for both sides argue about the case but do not introduce new evidence. Judges are quasi-randomly assigned in both rounds.

2.2.1. Model

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

$$r_i > \varepsilon_{ijs}(X_{ijs}, W_{ijs}) = \gamma_{js} + X_{ijs}\beta_s + \widetilde{\varepsilon}_{ijs}(W_{ijs}) \tag{2.1}$$

where $\widetilde{\varepsilon}_{ijs} \sim G_{js}$ and $\exists x_{ijs} \in X_{ijs}$ s.t. $x_{ijs} \notin W_{ijs}$. Judge *leniency* is captured by γ_{js}; high levels of γ_{js} mean that fewer claimants are approved. This threshold can be adjusted by X_{ijs}.[7]

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

I operationalize consistency by comparing judges with the same approval rate. In the first round, the approval rate is

$$P[r_i > \varepsilon_{j1}(X_{ijs}, W_{ijs})] = \int G_{j1}(r_i - \gamma_{j1} - X_{ij1}\beta_1) f_r dr \tag{2.2}$$

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

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

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

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

In words, this definition is straightforward: a consistent judge is more likely to approve high-quality claimants, and less likely to approve low-quality claimants. I assume that for any

[7] It is mathematically equivalent if X_{ijs} shifts the distribution of r_i without otherwise affecting the distribution and the same X_{ijs} is included in each round, but conceptually difficult to imagine what X_{ijs} might do this.

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

Leniency and consistency are related to the fairness of the court in a simple way. A perfectly fair court would have all judges use the same threshold γ_{js}, and be perfectly consistent in their decisions (e.g. $\widetilde{\varepsilon}_{js} = 0$). Thus, any individual would be always approved by the court, or always rejected. Cross-judge variation in γ_{js} increases unfairness by increasing the variation in individual's outcomes generated by judge assignment. Similarly, inconsistency induces randomness in the outcome for each individual; unfairness is monotonically increasing in inconsistency.

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

$$P[r_i > \varepsilon_{j1}(X_{ij1}) \cap r_i > \varepsilon_{k2}(X_{ik2})] = \int G_{j1}(r_i - X_{ij1}\beta_1 - \gamma_{j1}) G_{k2}(r_i - X_{ik2}\beta_2 - \gamma_{k2}) f_r dr \tag{2.3}$$

2.2.2. Identification

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

2.2.2.1. Judge-assignment identification. Take two judges with the same first-round approval rate, A and B. Then, a higher share of the more consistent judge's claimants will be ultimately

approved by a common second-round judge, C. Suppose that judge A is more consistent. Then, abstracting away from covariates X_{ij1} and substituting $\widetilde{G}_{js}(r) = G_{js}(r-\gamma)$ for clarity, this can be seen by noting that:

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

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

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

Identification of second-round consistency follows a slightly different route, because we do not have a third round to use as a check. Instead, I attain comparability from a *non-limiting*

judge. Suppose we are trying to determine which second-round judge, A or B, is more consistent. I assume that there is a known first-round judge D that approves nearly anyone, and a known first-round comparison judge C. Formally, I require that $\widetilde{G}_{C1}(\cdot)/\widetilde{G}_{D1}(\cdot)$ is monotonically increasing wherever $\widetilde{G}_{A2}(\cdot) \neq \widetilde{G}_{B2}(\cdot)$. This is trivially satisfied when $G_{D1}(\cdot) = 1$ (judge D literally approves everyone), and can be satisfied when judge D is fairly consistent and has a very low threshold γ relative to judge C.

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

(2.5)

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

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

2.2.2.2. Regressors and identification. The between-judge comparisons that I discuss in the previous section are local; they measure relative consistency for judges with similar approval rates. To compare judges who approve different shares of claimants and to identify the scale of judge errors $\widetilde{\varepsilon}_{ijs}$ without resorting to functional form assumptions, additional large-support continuous regressors are required. These regressors affect judge thresholds γ_{js} but do not otherwise affect errors. In a nonparametric sense, they are used as special regressors to identify the distribution of the composite error (ie, $\widetilde{\varepsilon}_{ijs} - r_i$) for each round. I then assume that at least one component of β_s is the same between rounds, tying down the relative size of the composite errors and identifying the distribution of r_i separately from $\widetilde{\varepsilon}_{ijs}$. In a parametric model, instruments are not strictly necessary (the model is mechanically identified), but provide a source of identification beyond functional form and judge randomization. A full proof of identification is in Chen et al. (2000); which I reframe in terms of my model in Appendix Section A.2.1.

2.2.3. Interpretation

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

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

$$\widetilde{\varepsilon}_{ij1} = u_{ij} + e_{ij1} \tag{2.6}$$

The first component, u_{ij}, represents permanent disagreements, or inter-rater reliability. As I discuss in Section 3.3, refugee appeals at the Federal Court are assessed along both substantive (is this person really a refugee?) and procedural (did the government properly make its initial decision?) lines. In other words, one judge might always reject a claimant who has a strong procedural case and a weak substantive one, while a different judge who weighs substantive considerations more heavily might always approve him. Inter-rater differences might also arise from differential bias along racial or gender lines — if judge A approved only women and judge B only men, they would disagree on the decision for each case even if the pool of claimants was 50% each gender. u_{ij} is therefore a measure of how a judge's weighting of different facets of a case differs from the consensus.

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

A natural question concerns the relative size of r_i, u_{ij} and e_{ij1}. My model identifies the relative variance of r_i versus the composite error $u_{ij} + e_{ij1}$, but does not directly estimate the size of u_{ij} versus e_{ij1}. However, the strength of the *additional* predictive power of judge identity

has strong implications for the relative size of the errors. Suppose that the composite error was mostly inter-rater differences. Then, one would expect that some pairs of judges would both value the same type of cases (for example, cases that were particularly strong on the procedural merits, or involved an Asian claimant). By definition, the probability of approval in the second round for a claimant with quality r_i and judges j and k (and suppressing extra regressors) is

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

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

$$\mathbb{1}[\text{Approval by } j | \text{Approval by } k] = \beta P_{jk}(0) + \nu_{jk} + u_{ijk} \tag{2.7}$$

where $P_{jk}(0)$ is calculated from the model.[8] Under the null of no correlations in errors between judge pairs, the judge-pair fixed effects ν_{jk} should be jointly insignificant. This is a joint test of *all* the reasons pairs of judges could disproportionately agree or disagree but failure to reject suggests that inter-rater differences are not large. This is turn suggests that test-retest

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

errors $\check{\varepsilon}_{ijs}$ are larger than inter-rater differences $\bar{\varepsilon}_{ij}$. I implement this test in Section 2.4.6, and fail to reject the null of no judge-pair effects.

2.2.3.2. 1st versus 2nd round errors. The decomposition of errors into inter-rater and test-retest components is complicated in the second stage by the possibility that the judges gain additional information about the case in the second-round hearing. As I discuss in Section 2.3.2, there is a full hearing in the second round (in the first they just review documents), and judges may learn things about the case that color their views of its strength. This is conceptually distinct from both observational errors e_{ij2} and inter-rater disagreements u_{ij} in that this new information could reflect information about the true merits of the case. In the second round the error can be decomposed

$$\widetilde{\varepsilon}_{ij2} = u_{ij} + e_{ij2} + \mathscr{I}_{ij2} \tag{2.8}$$

where $\mathscr{I}_{ij2}$ is an information shock. This shock should be thought of as realized information that if explained (e.g., written down in an opinion) would shift the cross-judge consensus on case strength. The subscript j reflects that some judges may be better at finding this information, and thus have a wider distribution of $\mathscr{I}_{ij2}$. The interpretation of the other components is analogous to the first round: u_{ij} is a judge-level bias term that reflects judge-level tendencies to accept certain arguments or types of cases, and e_{ij2} is observational error.

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

perceptive judge, so the relationship between judge quality and second-round consistency could go in either direction. For this reason, I focus most of my discussion on $\widetilde{\varepsilon}_{ij1}$.

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

$$\int\int\int \Big\{ \mathbb{1}[r_i > \varepsilon_{ij1}]\mathbb{1}[r_i < \varepsilon_{ik1}] + \mathbb{1}[r_i < \varepsilon_{ij1}]\mathbb{1}[r_i > \varepsilon_{ik1}] \Big\} f_r \, d\varepsilon_{ij1} \, d\varepsilon_{ik1} \, dr \tag{2.9}$$

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

As I will show in Section 2.4.6, inconsistency seems to be driven mostly by observational errors rather than ideology. I interpret this to mean that $\widetilde{\varepsilon}_{ij1}$ is likely to be uncorrelated across judges in the same round. In that case, Equation 2.9 can be used to calculate the disagreement rate under the assumption that $\widetilde{\varepsilon}_{ij1} \perp \widetilde{\varepsilon}_{ik1}$. I refer to this as *uncorrelated disagreement.*

Alternatively, I bound the size of the disagreement and find the minimum level of disagreement for any joint distribution of $\widetilde{\varepsilon}_{ij1}$ and $\widetilde{\varepsilon}_{ik1}$. For each r_i, a pair of judges disagrees with at least probability $|G_{j1}(r_i - \gamma_{j1}) - G_{k1}(r_i - \gamma_{k1})|$. Total disagreement can then be bounded by

integrating over the distribution of case strength r_i. By construction, *bounded disagreement* is a conservative measure of disagreement, but will be larger than zero whenever judges vary in the distribution of their observational error.

2.3. Institutional Background and Data

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

2.3.1. Immigration and Refugee Board

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

origin of the claimant or the stated reason for the claim. The Members are political appointees rather than long-term, professional bureaucrats.

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

Claimants who have been rejected for refugee status may apply to the Federal Court for judicial review. Approximately 65% of denied claimants file an appeal, which allows most claimants to stay in Canada until the Federal Court makes its final decision.[9]

IRB procedures for making refugee determinations were broadly consistent from 1995 until December 15, 2012, when an administrative appeal division partially supplanted the review work of the Federal Court (Grant and Rehaag, 2015). The only major policy change in this period concerned the composition of the IRB panel that made the decision. For refugee claims

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

submitted before June 28, 2002, standard procedure was for the case to be heard by a two-Member panel. If either member recommended approval, refugee status would be granted. Upon consent of the claimant, the case could be heard by a single Member, and by 2002 this practice was common (Dauvergne, 2003). However, the claimant often knew which Member would be making the decision if they agreed to a single-Member panel. Ostensibly they would be less likely to let the decision on their refugee status be made by a Member with a low approval rate, meaning that they had some ability to pick who would decide their refugee status. After the implementation of the Immigration and Refugee Protection Act (IRPA) in 2002, all cases were heard by a single Member. This is important because it suggests that the distribution of case strenth for the rejected claimants who appeal to the Federal Court changed after IRPA came into affect; more high-quality claimants may have been rejected, skewing the distribution further to the right. To allow for this possibility, I allow for the distribution of case strength r_i to vary before and after IRPA. More details are in Section 2.3.5.

2.3.2. Federal Court responsibilities and protocol

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

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

The first-round judge makes her decision after reviewing written records from the IRB decision and briefs written by the lawyers for the claimant (arguing for a second-round hearing) and the government (arguing against). If they decide against judicial review, the claim is rejected and the claimant is usually deported.[10] If the petition for leave is approved, the case goes to a full judicial review (JR) hearing. The judge for the second-round hearing is also quasi-randomly assigned, so usually the second-round judge is someone different. Regardless of the first-stage outcome, the first-round judge does not provide a written explanation for her decision. This may be one reason to expect that second-round decisions will frequently be inconsistent with the first-round approval. It could also contribute to inter-rater inconsistency for first-round judges, since the dearth of written precedent makes it difficult for judges to learn about how their colleagues have ruled on similar cases. Finally, it might lead to a wider spread in standards (in my model, thresholds γ_j).

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

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

In both rounds, judges are not tasked with determining whether the IRB Member made the right decision. Under Canadian law, judges must show deference to administrative decisions. This means that instead of determining whether the "correct" ruling was made, the judge must simply decide whether the government's initial decision was "reasonable" (Rehaag, 2012).[11]

The Federal Court reviews IRB decisions on both substantive and procedural grounds, although the reasonableness standard means that the bar for overturning the decision is high. A substantive ground on which a judge might reverse an IRB decision would be if the Member had ignored credible evidence that a claimant had been tortured. In contrast, procedural reasonableness requires that the Member collect adequate testimony from the claimant. A judge would be expected to rule in favor of the claimant if there were large procedural violations, even when they believe that the claimant does not actually qualify for refugee status. However, the precise extent to which judges are supposed to weigh substantive and procedural factors is unclear, and it is natural to expect that different judges would differentially consider different aspects of the case. The extent to which this is true is one of the main factors that determines the size of inter-rater inconsistency.

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

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

Judge assignment works similarly in both stages. For the first stage, judges are assigned to cases using a pre-set schedule; in each office the judges rotate through "leave duty." When enough cases have accrued the court gives the leave duty judge all the outstanding files (usually on Monday), and they are responsible for disposing of all of them. There is no review of the cases before they are given to the judge, and the leave duty schedule is not public. Previous research claims that this assignment is as good as random (Rehaag, 2007); in Section 3.5 I show that judge leniency is uncorrelated with case or claimant characteristics predictive of success. In the second stage the assignment process is similar; cases are divided between judges who are available for refugee work without review of the contents. A computer program slots hearings into the available times in the judges schedules. Occasionally the same judge will be assigned to a case for both stages by chance. This is potentially important because it implies that between-round errors may be correlated when the same judge is making the decision. Interestingly, this could arise either from inter-rater differences (if a particular claimant has a strong substantive case but a weak procedural one, she would be more likely to succeed in both rounds if she was assigned a judge who heavily weighted substantive aspects) or test-retest errors (a judge might remember the claimant from making the decision in the first round, and so could make the same observational error). I explicitly allow for this by estimating an additional parameter for the correlation between judge errors $\widetilde{\varepsilon}_{ijs}$ in the two rounds whenever the same judge is making the decision in both rounds; more discussion is in Section 2.3.5.

2.3.3. Reform to selection of Federal Court justices

Federal Court justices are appointed by the Minister of Justice. Appointments are until the mandatory retirement age of 75.[12] For most of Canadian history the Minister has had nearly unfettered discretion over appointments and has used this power to reward "active supporters of the party in power" (McKelvey, 1985). The only check on the government was a committee of the Canadian Bar Association that offered non-binding advice on the suitability of candidates.

A major reform in 1988 reduced the discretion of the government in making appointments. The reform created province-level judical advisory councils (JACs) to pre-screen applicants before they could go to the Minister for possible selection. The committees were made up of one member of the provincial Law Society, one member of the provincial branch of the bar association, one representative of the provincial chief justice, one representative of the provincial attorney general, and three representatives of the Minister of Justice. The JACs rated each candidate as "highly recommended," "recommended," or "not recommended," and the government could pick judges only from the pool of recommended and highly recommended candidates. The standards concorded well with a lay understanding of what makes a good judge: "'professional competence and experience' (such as proficiency in the law, awareness of racial and gender issues); 'personal characteristics' (ethical standards, fairness, tolerance); and 'potential impediments to appointment' (drug or alcohol dependency, health, financial difficulties)" (Hausegger et al., 2010). Crucially, the direct representatives of the Minister were a minority on the committee, making it difficult to push through wholly unqualified candidates.[13] The standards had

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

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

bite; only about 40% of candidates were recommended or highly recommended. Although the government could ask a JAC to reconsider a candidate's rating, the reform seems to have reduced the level of patronage. Russell and Ziegel (1991) report that before 1988 at least 47% of appointed judges had some involvement with the ruling Conservative party.[14] Their data comes from reports by surveyed respondents in the legal progression, and so estimates are likely biased down. Though data on post-reform connections to the ruling party are not exactly comparable, Hausegger et al. (2010) search through administrative records and find that after the reform only 30% of newly-appointed judges had donated to the party in power in the five years before their appointment. This is consistent with the new system reducing the number of unqualified party supporters being appointed to the bench, and suggests that the overall consistency of the courts may have improved as a result. I will test this hypothesis in Section 2.4.9, and find evidence that consistency did improve after the reform.

2.3.4. Data

My main data come from Federal Court case reports available on their website.[15] I parsed the data and verified it against a smaller subset professionally transcribed by Rehaag (2012). I use all cases since 1995 that were filed at the IRB before the implementation of the Immigration and Refugee Protection Act (IRPA) on June 28, 2012 (as discussed in Section 2.3.1, IRPA created a Refugee Appeal Division within the IRB, substantially changing the number and type of refugee appeals at the Federal Court). I also require that the appeal at the Federal Court was filed before the end of 2012 to ensure that there was enough time for all cases to be disposed of.

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

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

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

Using the first name of the claimant, I infer gender using British Columbia and Social Security Administration birth records that contain both first name and gender. To collect information on the country of origin of the claimant, I link the court records to the subset of available IRB case files.[16] These data contain the name of the IRB Member who made the initial determination, and in some cases the country of origin and gender of the claimant. I also use the commercial service Onomap to predict country of origin for each claimant, which I collapse to continent dummies.

For each judge, I collected information on the date and party of appointment. Appendix Table A8 contains summary statistics for the judges. 25% are female, and their dates of appointment range from 1982 to 2010. Since the Liberals held power for most of this time period, 72% of judges are Liberal appointees.[17] The average judge has 6.5 years of experience with a maximum of 28.

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

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

I exclude cases that were not perfected[18] or were unopposed by the government. I include only judges who decided cases in both the first and second round to improve comparability between the estimates of first- and second-round judge behavior.

2.3.5. Estimation

Fully nonparametric identification as outlined in Section 2.2.2 requires special regressors with large support conditional on judge assignment. This is a very high bar, and one that is not met by my empirical application. Instead, I parameterize the distributions of case strength r_i and judge errors $\widetilde{\varepsilon}_{ij1}$ and $\widetilde{\varepsilon}_{ij2}$, generating tractable analytic expressions for approval probabilities and allowing rapid estimation of the entire model by maximum likelihood. I begin by assuming that $\widetilde{\varepsilon}_{ijs}$ is mean-zero and normally distributed with standard deviation σ_{js} to be estimated as the measure of judge inconsistency. Larger σ_{js} corresponds to more inconsistency, ie a wider distribution of judge errors $\widetilde{\varepsilon}_{ijs}$. I additionally allow regressors W_{ijs} to affect errors, so

$$\widetilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2), \quad \sigma_{js}(W_{ijs}) = e^{\widetilde{\sigma}_{js} + W_{ijs}\psi_s} \tag{2.10}$$

As discussed in Section 2.3.2, the distribution of unobserved case strength for claimants at the Federal Court is likely right-skewed, since it is the distribution of individuals who were denied refugee status by government decision-makers. A relatively small number of high-case strength refugees are likely *not* initially granted status by the government. To reflect this intuition I therefore assume that the distribution is single-tailed; r_i is exponential-Pareto distributed

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

(I show the distribution of r_i relative to the estimated parameters in Figure 4.16). I allow flexibility in the distribution of r_i across two dimensions. First, cases filed at different offices vary considerably in strength. This is partially because office is correlated with country of origin, but more closely related to varying levels of legal aid funding. Second, the government made changes to the decision process for initial refugee applications in 2002 (see Section 2.3.1 for more details). Combining these, I fix the distribution of r_i to have a scale and shape parameter of 1 for the largest office (Toronto) before 2002, and then separately estimate the scale and shape parameters for each office before and after 2002. In practice, I find almost no difference in the distribution of case quality between these time periods, suggesting that the institutional changes had little affect on which claimants were approved. However, the between-office variation is considerable.

In Appendix Section A.2.2, I show that the probability of first-round approval with exponential-Pareto location parameter is x_m and scale parameter α is

$$P(r > \widetilde{\varepsilon}_{j1}(X_{ij1}, W_{ij1})) = \Phi\left[\frac{\ln(x_m) - \gamma_{j1} - X_{ij1}\beta_1}{\sigma_{j1}(W_{ij1})}\right] + e^{\alpha(\ln(x_m)-\gamma_{j1}-X_{ij1}\beta_1)+\frac{\alpha^2\sigma_{j1}(W_{ij1})^2}{2}}\left[1-\Phi\left(\frac{\ln(x_m) - \gamma_{j1} - X_{ij1}\beta_1}{\sigma_{j1}(W_{ij1})} + \alpha\sigma_{j1}(W_{ij1})\right)\right] \tag{2.11}$$

which can be calculated without resorting to computationally expensive numerical integration. Joint probabilities for first and second round approval are similar in spirit but more complicated. Since occasionally the same judge is assigned to the first- and second-round decision, I allow $\widetilde{\varepsilon}_{ij1}$ and $\widetilde{\varepsilon}_{ik2}$ to be correlated (with the correlation estimated as an additional parameter) whenever $j = k$. The full derivation and presentation is available in Appendix Section A.2.2.

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

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

I instead use the timing of the decisions during the week, and of the second-round hearing during the day, as regressors. Danziger et al. (2011) argue that when decision-makers make many decisions in a row, they become fatigued and are more likely to pick the default option. They study parole decisions in Israel and find that rejections become more likely just before lunch and revert to baseline levels immediately after the break. I follow them and use a dummy for the noon hearing (in the second round only) as a regressor.[19]

Second, I exploit the fact that judges make decisions on refugee cases only irregularly (non-refugee cases occupy much of their time). In the first stage, they are given a tranche of case files at the start of the week and work on them until they have made all the decisions. One might expect that judges would become more fatigued as the week goes on. Ideally, I would define a decision as having been made late if it was made on Tuesday or later, after the first full day of decisions. However, judges may endogeneously change the order in which they make decisions as a function of case characteristics (concretely, they may delay the difficult decisions), which

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

would violate the exclusion restriction. I therefore include a dummy for the last submission before the judge's decision happening on Wednesday or later as a regressor — empirically, this predicts the leave decision is likely to be made Tuesday or later. Similarly, for second-round cases the scheduling is done by court staff without knowledge of the case characteristics. I define the end-of-week regressor in the second round as the case being heard on Wednesday or later (Monday hearings are rare, so Wednesday is usually the second day of hearings). I assume that the effect of a late-week hearing is the same in both the first and second round.

Identification using regressors leans on the assumption that variation in X_{ijs} does not affect errors. One might reasonably wonder whether this is true for the timing regressors. I address this issue in two ways. First, I estimate versions of the model where the excluded components of X_{ijs} are added in turn to W_{ijs} — I allow the regressors to directly affect the size of inconsistency. Since there are two regressors in the second round (dummies for the end of the week and a lunchtime hearing), I can then test whether $\sigma_{j2}(W_{ij2})$ varies with the regressors. In Section 2.4.5, I conduct these tests and find that the effect of the regressors on the error distribution is small and statistically insignificant.

Second, I estimate the model without relying on regressors. This version of the model leans more heavily on functional form assumptions to compare consistency for judges with different approval rates, but is a valuable robustness check. In Appendix Section A.2.7, I present estimates for this model and show that all the results are qualitatively similar.

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

2.4. Results

2.4.1. Randomization tests

In Table 4.6 I explore whether the cases are assigned quasi-randomly to judges. One implication of quasi-random assignment is that judge characteristics should be unrelated to case characteristics. To test this, for each round I regress claimant characteristics on judge-level mean approval rates in that round, controlling for office X pre-2002 fixed effects to account for office and time variation. I also regress the characteristics on judge fixed effects and report the F-stat and p-value for the joint test of the judge fixed effects.

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

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

Claimant demographics come from the IRB case files, which are only available for about 85% of cases. In Appendix Table A9, I display similar regressions for the entire sample, substituting gender and continent of origin imputed from claimant name as dependent regressors. Judge leniency has some predictive power for imputed continent of origin, but not in a way that is correlated with predicted approval.

2.4.2. Reduced form judge behavior

Federal Court judges are obliged to show deference to the government's initial determination of refugee status. Perhaps because of this, approval rates in the first round are low, at only 14%. There is a large amount of cross-judge heterogeneity: the histogram in Panel A of Figure 4.12 shows that four judges approved less than 5% of cases, while one judge approved 70% (after this judge, the next highest rate is 28%).

In the second stage the approval rate is much higher, at 44%. Similarly to the first round, there is a large amount of dispersion in approval rates, from 13% to 87%. The dramatic improvement in the success rate in the second round suggests that first-round judges are effective to some degree in terms of choosing claimants who can, in the language of the Court, make an "arguable case" in the second round. In terms of the structural model, this implies that variation in refugee quality r_i is substantial and (at least partially) commonly observed by judges. More evidence in favor of a latent factor r_i can be seen in Panel C of Figure 4.12, which shows that there is a high correlation (0.56) between the first- and second-round approval rates for the same judge.

To the extent that there is a common case strength factor observed by judges, claimants approved in the first round by strict judges should fare better in the second round than those

approved by lenient judges. Table 4.7 conducts this analysis, regressing second-round approval on the exclusive mean approval rate for the first-round judge. Moving across the columns, I include no other controls, the mean approval rate of the second-round judge, and second-round judge FEs. The results are similar; *having been approved* by a 10 percentage point more lenient first-round judge gives you a 2.6-3.2 percentage point lower chance of being approved in the second round. The straightforward interpretation is that individuals who were approved by a more lenient judge in the first stage have, on average, a weaker case in the eyes of the second-round judges. This again suggests that there is a refugee quality factor r_i that is commonly observed by the judges up to some observational error.

A more structural way to demonstrate the existence of a commonly-observed quality factor is to estimate the marginal treatment effect (MTE) of first-round approval on ultimate approval, instrumenting for first-round approval with judge assignment. I include this graph in Appendix Section A.2.4, where I confirm that individuals marginally approved by more lenient judges in the first round are less likely to be approved in the second round.[20] The relationship is relatively weak and foreshadows the structural result that first-round judges have trouble selecting the claimants that will be successful in the second round; over the main mass of first-round approval rates the MTE declines from about 0.45 to 0.30.

Figure 4.12 demonstrates that there is substantial variation in judge propensity to approve refugee claims — in the language of the model, that there is variation in γ_{js}. For evidence on variation in σ_{js}, judges' ability to pick the highest-quality claimants, I turn to Figure 4.14. One source of identifying variation in the structural model comes from how often a first-stage judge's approved claimants are approved by a different judge in the next round. First-round judges with

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

more ultimately successful claimants, the model implies, are better at picking high-r_i cases. Figure 4.14 displays reduced form evidence on the size of the variation in this ability. For each first-round judge, I take the mean approval rate of her approved claimants in the second round. Panel A displays the histogram: a 10^{th} percentile judge has 37% of his claimants ultimately approved; a 90^{th} percentile judge 56%. Although it stands to reason that claimants approved by lenient judges in the first round would have a lower second-round success rate, the histogram changes little when I residualize out first-round approval rates and second-round judge approval rates (Panel B). In Panel C, I similarly calculate the second-round approval rates for claimants approved by each judge and plot them against the judges' first-round approval rates, residualizing out the second-round judge approval rate and shrinking the estimates toward the grand mean via Empirical Bayes to account for small cell sizes. The regression coefficient is negative and significant, but there is a large degree of cross-judge dispersion in second-round approval for each first-round approval average — the subsequent approval rate for judges approving about 15% of first-round applicants ranges from 38 to 48%. In other words, there is a lot of variation in the ability of judges to pick claimants who will be approved in the second round, even holding constant the share of claimants they approve.

2.4.3. Structural results

My baseline model includes as regressors X_{ijs} a dummy for a late-week decision and a dummy for whether the second-round hearing was heard at lunch. Thresholds γ_{js} vary by judge and round, and inconsistency σ_{js} varies by judge-round but not by any covariates (ie, W_{ijs} is empty). Case strength r_i is distributed as an exponential-Pareto, with different parameters in each office

of case origination and before and after 2002, when there were changes to how the government made initial refugee determinations. I discuss these modeling decisions in detail in Section 2.3.5.

Figure 4.16 plots the distribution of judge-round thresholds γ_{js} and inconsistency σ_{js}. The red dotted lines are the raw coefficients. Because of estimation error the distribution of the raw coefficients is slightly too wide; in blue I plot the distribution of the underlying coefficients after deconvolving out the measurement error using the method of Delaigle et al. (2008). The coefficients are precisely estimated, so for most of the estimates this does not make a difference. For comparison I plot the distribution of case quality r_i in black.

In Panel A, the distribution of γ_{j1} is large relative to the distribution of refugee strength r_i, plotted in black. In Panel C, the distribution of second-round thresholds γ_{j2} is narrower and slightly smaller on average.[21] I interpret this as reflecting the larger amount of precedent available to judges in the second round relative to the first round. In the first round, no decisions are written up, making it difficult for judges to learn about the decisions other judges have made in similar situations. Conversely, nearly all second-round decisions are published as precedent. The correlation between round-specific γ_{js}'s is relatively high (0.46).

Panel B of Figure 4.16 shows the round-specific distributions of σ_{js}. The standard deviation of the first-round judge errors for the median judge is 1.08, which is large relative to the standard deviation of 1 for underlying case quality r_i. To contextualize the size of the inconsistency, I match pairs of first-round judges with approval rates within one percentage point. Using the two methods of quantifying disagreement I discuss in subsubsection 2.2.3.3, I bound average disagreement at 3.6% of all cases. Using the more realistic assumption that cross-judge errors are

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

uncorrelated, I calculate that the average pair of judges with the same first-round approval rate disagrees on 13.2% of all cases. This is remarkably close to the estimated disagreement rate of 14.6% for first- and second-round judges if second-round judges saw all cases rather than only approved ones (for this comparison, the cross-judge errors are uncorrelated by construction).

How big is a disagreement rate of 13.2%? For *all* pairs of judges, the average disagreement rate is 23.1%, suggesting that inconsistency for similarly-severe judges is a larger contributor to how the justice system delivers different verdicts for the same claimant than cross-judge variation in severity. Alternatively, if all judges were perfectly consistent, cross-judge variation in approval rates alone implies that the average pair of judges disagrees on 8% of cases.

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

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

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

I conduct this exercise, and estimate that this measure of disagreement is bounded at 15.2%. Uncorrelated disagreement is a shocking 57.6%.

The presence of inconsistency has important implications for examiner-assignment research designs. This identification strategy uses random or quasi-random assignment of decision-makers to cases to generate random variation in a treatment, then studies the effect of treatment on outcomes. Prominent examples of the treatments studied include incarceration, patent receipt, and being placed in foster care (Bhuller et al., 2016; Gaulé, 2015; Doyle, 2008). The montonicity assumption in this context requires that all individuals are weakly more likely to be approved by a high-approval judge, and less likely to be approved by a low-approval judge (individuals for whom this does not hold are defiers in the Angrist et al. (1996) sense). The presence of inconsistency implies violations of monotonicity. In Appendix Section A.2.6 I examine the effect of inconsistency on IV and MTE estimates. Although this may be a setting where inconsistency is larger than other examiner-assignment contexts (suggesting large biases from inconsistency), the results are relatively reassuring. In a regression of second-round approval on first-round approval instrumented with judge assignment, even this relatively high level of inconsistency biases IV estimates by only 5%. More worryingly, inconsistency causes the estimated MTE to be considerably flatter than the true MTE.

In Panel D of Figure 4.16, the size of the distribution of second-round errors σ_2 is harder to interpret. Recall from subsubsection 2.2.3.2 that the second-round error may contain an informational component $\mathscr{I}_{ij2}$. If this is the case, then a larger σ_{j2} may reflect better information-gathering abilities in the second-round hearing. Some suggestive evidence that information-gathering is an important part of second-round errors is in Appendix Figure A8, a scatter of

judge-specific first- and second-round σ_{js}. The correlation is small and slightly negative (-0.043). It is reasonable to assume that the non-information components of first- and second-round errors are positively correlated, given that all the other measurable aspects of judge behavior (thresholds γ_j, overall approval rates) are correlated across rounds. This suggests there is a non-trivial information shock in the second round. The size of this shock is negatively correlated with the magnitude of first-round inconsistency — consistent first-round judges are also better at uncovering relevant information in the second-round hearing.[23,24] I take this as a further reason to focus on first-round consistency for the remainder of the paper.

Another interesting finding from Figure 4.16 is that most judge's estimated σ_{j2} is smaller than the σ_{j1} (though there is a long tail of large σ_{j1}). Why is this so striking? By construction, the information shock is orthogonal to other errors. This means that the distribution of judge errors $u_{ij} + e_{ij2}$ — which is necessarily smaller than $\widetilde{\varepsilon}_{ij2} = u_{ij} + e_{ij2} + \mathscr{I}_{ij2}$ after netting out informational shocks $\mathscr{I}_{ij2}$ — is smaller in the second round than in the first. This likely reflects that second-round judges take more time to make the decision, think more deeply, and usually write out a full decision explaining their reasoning.

First-round judges do a poor job of predicting which claimants will be successful in the second round; for most claimants the probability of second-round approval is higher than their probability of first-round approval.. Figure 4.18 shows how this works. For each r_i I calculate the first-round approval probability and the second-round approval probability conditional on first-round approval. I then plot them against each other. The figure shows that the median

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

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

claimant has a 6% chance of first-round approval, but conditional on approval has a 12% chance of approval in the second round. This does not reflect selection, which is accounted for by conditioning on r. Instead, it shows that second-round decisions are unpredictable from the perspective of the first round, and even claimants with a relatively low quality factor r_i are sometimes approved in the second round. Furthermore, no one has a very high chance of approval: the 95$^{\text{th}}$ percentile claimant has only a 62% chance of first-round approval and a 66% chance of second-round approval — a 41% chance of overall success at the Federal Court. The first-round approval rate is 14%; if the highest-r_i 14% was selected in the first round the overall approval rate would only climb to 8.5% (14% $\times$ 0.58) from its current value of 6%.

Integrating over the entire distribution of r, I find that on average 19.4% of claimants would be approved in a second-round hearing if they were approved in the first round. This is in contrast to the 6% of claimants who are approved under the current system. Although it may be surprising that the number of refugee appeals granted by the Federal Court would triple under this alternative decision-making process, the result is foreshadowed by the scatter plot of first-round approval rates by judge against the approval rates for that judges' approved claimants in the second round found in Figure 4.14. The most lenient judge approved 66.7% of claimants in the first round after partialling out the office and date of origin, and of those 27.4% were approved in the second round (partialling out the second-round judge, office and timing), bounding the overall approval rate in the absence of a first round at 18.3% (0.667 $\times$ 0.274). In terms of the policy implication, an important caveat is that judges might change their second-round behavior in unpredictable ways if all cases were automatically approved in the first round — one first-order effect would be that each judge would have 7 times more second-round cases to hear, which is likely not possible. This finding should therefore be interpreted more as a description

of how well first-round judges can select the claimants that will be successful in the second round than a prediction of what would happen under a one-round system.

2.4.4. Relationship between structural parameters and reduced form statistics

In this section, I show how reduced-form moments translate into the structural parameters. In the model, the principal determinant of approval rates is judge thresholds γ. In Panel A of Figure 4.19, I vary the γ_{j1}'s from their estimated values by adding a common shifter to each γ_{j1}. As they change, the estimated approval rate moves away from the observed value of 14%. Reassuringly, the change is monotonic and steep.

Recall from subsubsection 2.2.2.1 that a main source of identification of the overall size of first-round judge inconsistency σ_{j1} is whether the approved claimants are subsequently approved in the second round. In Panel B, I adjust σ_{j1} away from its estimated values by multiplying each coefficient by a common factor. As inconsistency increases (σ_{j1} gets smaller), this dramatically reduces the second-round approval rate.

Panels C and D hew closer to the judge-randomization intuition of subsubsection 2.2.2.1. Comparing two judges with the same first-round approval rate, the more consistent judge will have the higher approval rate in the second round for her approved claimants. In Panel C, I match judges with first-round approval rates within 1 percentage point of each other, then plot the difference in second-round approval rates against the difference in estimated σ_{j1}. As expected, judges with a higher second-round approval rate than their matched colleague have a lower estimated σ_{j1}, or in terms of the model are more consistent.

For second-round judges, identification relies on matching pairs of second-round judges with similar approval rates conditional on first-round approval by a very lenient first-round

judge. Equation 2.5 shows that second-round approval rates conditional on first-round approval by a different, less lenient judge will be higher for the more consistent second-round judge. Panel D of Figure 4.19 shows how the estimated model reflects this logic. Fortunately, my data contain one judge who approves 70% of first-round claimants, while the next-most-lenient judge approves only 28%. I match second-round judges by approval rates conditional on first-round approval by the outlier judge, taking all pairs with approval rates within 5 percentage points. In Panel D I display a binned scatter plot of the within-pair difference in estimated second-round inconsistency σ_{j2} and the difference in second-round approval rates conditional on first-round approval by all other judges. In line with the identification intuition of Equation 2.5, the larger the difference in approval rates, the larger the difference in estimated σ_{j2}. Higher approval rates under the comparison judges correspond to higher consistency (lower σ_{j2}).

2.4.5. Decision timing as regressors

Identification requires that the case timing regressors affect judge thresholds γ_{js} but are not correlated with judge errors $\widetilde{\varepsilon}_{ijs}$ or case strength r_i. I explore whether the regressors are uncorrelated with case strength in Table 4.8. Because case strength is unobserved, I predict first- and second-round approval from country of origin and gender of the claimant, then test whether this omnibus measure of r_i can be predicted by the regressors. The coefficients in Columns 1 and 2 are small and insignificant, suggesting that the timing of the cases is uncorrelated with case strength. In Columns 3 and 4, I show that the timing of the decision has a significant effect on both first-and second-round approval. This is important because it suggests that X_{ijs} sizably affects judge thresholds γ_{js}, and that the regressors make a substantive contribution to identification.

Another fear is that decision timing affects approval through changing judicial errors $\widetilde{\varepsilon}_{ijs}$ rather than case strength r_i or judge thresholds γ_{js}. If this were true, one possible implication is that the regressors would affect the distribution of errors. I test this directly in Appendix Table A10, where I include in turn the two decision timing regressors in W_{ijs}. In a nonparametric sense, the model is identified by an excluded regressor in each round that affects thresholds but not errors; since the noon-hearing regressor affects errors only in the second round these should be interpreted as tests on second-round identification. In line with the relevance tests in Table 4.8, both regressors have strong and statistically significant effects on γ_{js}. However, neither has a statistically significant affect on σ_{j2} — the log-log end-of-week coefficient is only 0.04 (SE=0.06) and the noon-hearing coefficient a very imprecisely estimated 0.37 (SE=1.29).

As an additional robustness check, in Appendix Section A.2.7 I estimate the model without regressors and find that all the main results are qualitatively unchanged, albeit slightly less precise. Because this version of the model is identified only using judge randomization and functional form, the fact that the results are similar suggests that the effect of X_{ijs} on judicial errors is small.

2.4.6. Ideology versus observational errors

In subsubsection 2.2.3.1, I discussed how judicial inconsistency can be decomposed into off-consensus ideological differences (inter-rater inconsistency) and pure observational errors (test-retest inconsistency). In this section, I provide some evidence on which factor is more important.

The Federal Court occasionally assigns the same judge to both the first- and second-round decision. I model this by allowing the observational errors $\widetilde{\varepsilon}_{ijs}$ to be correlated between rounds

whenever the judge is the same in both rounds. The correlation is estimated to be positive and quite large, at 0.32 (SE=0.04). In a reduced-form sense, this reflects second-round justices disproportionately approving claimants that they approved in the first round, above and beyond what would be expected by their overall first- and second-round approval rates. One could interpret such a high correlation as resulting from either inter-rater inconsistency (a judge disproportionately values the strengths of the claimants case) or a common misreading of the facts of the case in both rounds, which I refer to as test-retest inconsistency. However, the size of the correlation implies that at least one of these factors is large. In subsubsection 2.2.3.1 I describe a test of the relative size of these errors. If the correlation is caused by ideological inter-rater inconsistency, it is likely that pairs of judges share the same weighting of different aspects of cases. Then, in a regression of second-round approval on model-predicted likelihood of approval *and* judge-pair fixed effects, the fixed effects should add meaningful predictive power. One interpretation of this regression is as an overidentification test — the model is estimated under the assumption of no correlation between errors for different pairs of judges.

I implement the test in Table 4.9. The left two columns take order into account when constructing the judge pairs (ie, judge A then judge B is different from judge B then judge A), while the rightmost two columns ignore ordering. All specifications drop cases where the same judge made both the first- and second-round decisions.

Across columns, the coefficient on model estimates of approval probabilities is very close to one. However, in all specifications the F-stat for the joint test of judge-pair fixed effects is about 1, corresponding to p-values in the range of 0.30-0.60.[25] In other words, the judge-pair effects do not predict second-round approval beyond the model estimates. This suggests that

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

there is in fact no correlation between errors $\widetilde{\varepsilon}_{ijs}$ for judge pairs, and that ideological errors are small relative to observational errors. As a descriptive analysis, I calculate the Empirical Bayes judge-pair means.[26] This confirms the F-stat result: the standard deviation of the judge-pair means is only 0.004 relative to a mean approval rate of 0.44.

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

---

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

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

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

2.4.7. Judicial inconsistency, experience and workload

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

Table 4.10 presents models where I allow experience to enter the judge threshold γ_s (ie, in X_{ijs}) and the variance of the error, σ_{js} (ie, in W_{ijs}). I parameterize experience with an indicator for more than one year of experience and with indicators for more than 1, 5, and 10 years of experience. Column 1 shows that first-round log inconsistency σ_{j1} shrinks dramatically after the first year (0.77 log points), followed by smaller but still substantial decreases of 0.29 log points after five years and 0.54 log points after ten years. To put these numbers into context, I estimate that if all judges had less than one year of experience, pairs of judges with the same approval rate would disagree on 74% of approved cases when judge errors are uncorrelated. After one year of experience for all judges that declines to 54%; after a total of 5 years, 45%; and after 10 years 30%.[29,30] The pattern is similar when disagreement is measured as a share of all cases: judges with the same first-round approval rate disagree on 22% with no experience, then 17, 14, and 10% after 1, 5, and 10 years of experience.

This pattern of front-loaded gains to experience is similar to that observed in teachers, who see the most dramatic gains after the first year (Rivkin et al., 2005). In contrast to teachers, I see

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

[30] Interestingly, the decline in bounded disagreement is not monotone. For 0, 1, 5, and 10 years of experience for all judges, disagreement shares for approved cases are bounded at 9.2, 10.3, 9.6, and 7.5%.

further gains even after 10 years of experience, perhaps reflecting the more complicated nature of judging.

A potential implication of these high returns to experience is that more cases should be given to experienced judges. This will be true to the extent that judges improve as a function of years on the job, rather than directly from the number of cases they have made decisions on. The latter scenario is perhaps more likely at the Federal Court than elsewhere. Recall from Section 2.3.2 that first-round decisions are not written up as precedent. What's more, the second-round decision is not made for months (the median wait is 89 days), perhaps making it difficult for first-round judges to follow up on which specific cases have been approved in the second round. The only ways that judges can learn how their colleagues make decisions in the first round are indirect: inferring them from the types of cases they observe themselves in the second round, and discussing cases with colleagues. Both of these factors suggest that years of experience will be more relevant than number of cases see. Fortunately, we can take this distinction to the data. In Column 2 of Table 4.10, I allow years of experience *and* career number of cases to affect inconsistency. Analogously to experience, I implement the control for number of career cases with dummies, setting the thresholds at the 95^{th} percentile of the number of cases seen for judges with fewer than 1, 6, and 11 years of experience. Interestingly, the results are nearly the same: inconsistency declines by 0.69, 0.24, and 0.44 log points after 1, 5, and 10 years of experience, respectively. In contrast, the effect of number of cases on inconsistency is small and statistically insignificant, at effects (in log points) of 0.16 (SE=0.53), 0.066 (0.11), and 0.18 (0.67) after the career number-of-case thresholds corresponding to 1, 5, and 10 years of experience. This suggests that the Federal Court could indeed improve overall consistency by shifting some cases from inexperienced to experienced judges.

Another factor that may affect judicial consistency is workload. Higher caseloads may reduce the amount of time judges can spend on cases, or make them work longer hours. Chen et al. (2016) show that refugee judges in the US make more gambler's fallacy errors — judges being less likely to approve a claimant when they approved their last claimant — when caseloads are high, so one might expect to observe the same phenomenon here. To test this, I calculate monthly log workload as the number of leave cases a judge is assigned in a given month. Judges are also responsible for non-refugee cases, so this is an imperfect measure of workload. The model accounts for judge-specific consistency in σ_{js}, guaranteeing these estimates do *not* reflect time-invariant selection of more- or less-consistent judges into refugee work.[31] However, to the extent that judges with higher refugee caseloads may have lower non-refugee caseloads, these estimates are likely biased towards zero.

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

2.4.8. Judge inconsistency and expert opinion

The judge inconsistency parameters are related to readily-observable reduced form moments in the data, as well as with experience and workload in largely predictable ways. In this section, I explore whether they are also related to lawyer perceptions of judge ability. Higher degrees

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

of correlation between model-based measures and expert opinion serve as a validation of the model, and suggest that it could be used as a diagnostic tool.

To measure expert opinion, I conducted an email survey of refugee lawyers who have appeared in refugee hearings at the Federal Court. I asked respondents to rate the judges with whom they had personal experience along dimensions analogous to the parameters of the model: how lenient is the judge to claimants (corresponding to judge threshold γ_{js}), and how consistent and predictable is the judge (I reverse this scale so it corresponds to judge inconsistency σ_{js}). More details about the survey, including the question text and comparison of the respondents to the lawyer population, are in Appendix Section A.2.5.

Each response is on a five-point likert scale, which I normalize by the mean and standard deviation. Table 4.11 describes the relationship between model coefficients and the survey results. I model the relationship as

$$\widehat{C}_{j\ell} = \beta_0 + \beta_1 \text{Favorability}_{j\ell} + \beta_2 \text{Inconsistency}_{j\ell} + \eta_\ell + u_{j\ell} \tag{2.12}$$

where ℓ indexes lawyers and $\widehat{C}_j = \{\widehat{\gamma}_1, \widehat{\gamma}_2, \widehat{\sigma}_1\}$. I use model estimates that account for experience (which is highly predictive of behavior), and for each judge-respondent pair use the coefficient combinations reflecting experience at the time of their modal interaction. To account for estimation error in the model coefficients I use Hanushek's (1974) efficient estimator.[32] In Panel A, the dependent variable is the first-round $\widehat{\gamma}_1$. As expected, higher lawyer-reported favorability is associated with a lower threshold. The correlation is large but imprecise in the first

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

round; in the right-most preferred specification one SD higher favorability corresponds to a 0.19 lower γ_1, which is about 0.21 SD of the cross-judge distribution of γ_1. Panel B displays the relationship between γ_2 and the survey measures. The relationship is stronger than with γ_1; adding one SD of predicted favorability decreases γ_2 by 0.42, or 0.42 SDs of the judge distribution. This may be because second-round judge behavior is more salient than first-round behavior for lawyers, since they appear in front of the judge only in the second round.

Finally, Panel C shows the relationship between surveyed judge characteristics and σ_1. Survey inconsistency is positively related with the model estimate of inconsistency σ_1; across specifications one extra SD of inconsistency translates to between 0.16 to 0.22 higher σ_1, or about 0.1 SD of the cross-judge distribution. For comparability to Panels A and B, Panel C includes the same set of controls, but in Appendix Table A13 I show results slightly more in line with the identification intuition of subsubsection 2.2.2.1. I control for first-round approval rates, thus comparing σ_1 for judges with the same approval rate.[33] The results are reassuringly similar. In the right-most preferred specification, one additional SD of surveyed inconsistency corresponds to 0.13 higher σ_1. Across specifications, the model and the judge survey select the same judges as being more consistent, suggesting that the structural model is picking up true variation in judge ability to assess case strength and use common standards.[34]

[33] Analogously to how I define σ_1, I control for the judge approval rate for the level of experience when he had his median interaction with the judge. To maximize reach I surveyed lawyers who had appeared in front of judges for appeals of deportation and consular decisions involving refugees; this means I drop some observations with no inland refugee cases during that time period. The results are nearly identical if I control for whole-career judge approval rate.

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

2.4.9. Judge selection reform and judge consistency

In 1988, the government enacted an important reform to how it selects judges. The goal of the reform was to make it harder for the party in power to appoint unqualified party supporters. As I detail in Section 2.3.3, the limited evidence available suggests that the policy change reduced the number of new judges with ties to the ruling party. In this section, I provide evidence that the reform was also successful in reducing judge inconsistency σ_{j1}.

Table 4.12 presents a regression of $\widehat{\sigma}_{j1}$ on a dummy for whether the judge was appointed before the reform. I weight the regressions to account for estimation error in the dependent variable (Hanushek, 1974), and in my preferred specifications control for judge gender and party of appointment. Because the reform took place seven years before the start of my sample, the pre-reform judges are mechanically more experienced. More experienced judges are more consistent (have lower σ_{j1}), so this likely works against finding that the reform improved judge consistency.[35]

I show results for a baseline model that does not control for experience, and controlling for experience with categorical variables for more than 1, 5 and 10 years of experience (for approximate comparability, I adjust all coefficients to the median experience of 6 years). The first three columns show that average consistency improved by 0.34 after the reform, a 26% reduction (the estimate is just shy of statistical significance, with a p-value of 0.12. This does not appear to be related to the change in party that occurred just after the reform — Column 3 shows that party has no affect on large of statistically significant effect on consistency. In

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

Columns 4-6, the the effect is larger once the model properly accounts for differing experience among pre- and post-reform judges, with σ_{j1} dropping by 1.75 points (relative to a pre-reform mean of 2.2) for judges appointed after the reform.[36]

Both of these affects are large. The estimates from my preferred right-most model imply a pre-reform uncorrelated disagreement rate of 18.3% for pairs of judges with the same approval rate (recall the baseline estimate is 13.2%). If all judges were appointed post-reform, that drops to 6.6%.[37] The strength of the effect speaks to the size of the reform, which materially restricted the minister's options. Government data shows that the Judicial Advisory Councils approve only 40% of applicants; ostensibly some of the rejected candidates would otherwise have been appointed. Interestingly the effect is not driven by changes in judge leniency. In Appendix Table A14, I show that judges appointed after the reform approved a similar share of first-round claimants (an insignificant 5 percentage points more). More directly, in Appendix Table A15 I control for judge approval rates and find almost identical results, with an average σ_{j1} 1.72 smaller for judges appointed after the reform (SE=0.65).

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

[36]In Table A16, I estimate an identical table using second-round inconsistency σ_2 as the dependent variable. Because σ_2 may partially reflect skill at gathering information in hearings, the predicted effect is ambiguous. In all but one specification, I find statistically insignificant results. This suggests either that post-reform judges made fewer errors only in the first round (which is implausible), or that information acquisition is an important part of second-round errors.

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

consistent than female judges, though this difference is dwarfed by both the gains to experience and the post-reform effect.

2.4.10. Optimal judge allocation

The Court assigns cases to judges taking into account only their availability, not their behavior in previous cases. In this section, I show how the Court could optimize judge allocation to minimize caseload while maintaining the same standards.[38]

Second-round decisions are much more costly to the court than first-round decisions. Instead of reading documents from the IRB's initial determination, a second-round decision entails a full hearing in front of the opposing lawyers, time to prepare for the hearing and time to write the decision — about ten times as long as a first-round decision. The Court could minimize workload while approving the same number of total claimants by reducing the number of first-round acceptances and approving all second-round claimants. To some extent, they are already pursuing this strategy — as I discuss in Appendix Section A.2.4, the marginal claimants for most first-round judges have a relatively high chance of second-round approval (30-45%).

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

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

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

In Figure 4.21, I conduct exactly this maximization. I find that overall workload would be reduced by 17.5% (or 28,000 hours), amounting to savings of approximately $4.4 million in judge salaries alone over the study period. This counterfactual poicy would also save staff time and allow claimants to receive their ultimate decision faster. The figure demonstrates the second two kinds of savings, but not the first. To summarize how the re-assignment procedure works, I present histograms of the baseline judge coefficients by round as well as histograms that have been reweighted to reflect the distribution of coefficients after optimization. The average first-round threshold γ_{j1} for optimally-assigned judges is higher (Panel A), meaning that fewer cases will be approved in the first round but case strength conditional on first-round approval will be higher. Conversely, judges in the second round are much more lenient, as evinced by the lower thresholds γ_{j2} (Panel C). However, in Panel B and D the change in the overall distribution of consistency is much less dramatic — only the most inconsistent judge-rounds are eliminated. I interpret this to mean that judge thresholds γ_{js} are a stronger driver of who is selected, but that knowledge of judge-specific consistency has an important role to play in allowing the researcher to discipline the selection process.

The problem as I've described it takes the posterior distribution of r_i as the relevant measure of quality. Implicitly, this assumes that the second-round error is all judge error. As I discuss

in subsubsection 2.2.3.1, it may also reflect additional information gained in the second-round hearing. In that case, the relevant measure of quality is $r_i + \mathscr{I}_{ij2}$. I do not directly estimate the distribution of of the information shock $\mathscr{I}_{ij2}$, so cannot perfectly condition on the posterior (I estimate the distribution of inconsistency plus information shock, $u_{ij} + e_{ij2} + \mathscr{I}_{ij2}$). However, under the assumption that the second-round error is *all* information, I can minimize workload so that the posterior of $r_i + \mathscr{I}_{i2}$ first-order stochastically dominates the baseline distribution. Under this specification, the allocation of judges is highly correlated with the baseline optimization (0.58), and workload is reduced by 16% rather than 17.5%. It also satisfies the constraints of the baseline allocation, so a cautious planner could implement the second design and enjoy most of the gains of judge reallocation.

2.5. Conclusion

Much research has focused on non-relevant factors that affect judge behavior: the decision in the previous case (Chen et al., 2016), the outcome of a college football game (Eren and Mocan, 2016), or the timing of the hearing relative to lunch (Danziger et al., 2011). The existence of these phenomena suggests that the same defendant could be convicted by one judge and acquitted by another, even when both judges have the same overall incarceration rate. This behavior, which I call inconsistency, violates the fundamental ideal of fairness in the judicial system: that all cases should be decided according to the same standards. To some extent we already know that courts are unfair — cross-judge variation in incarceration rates implies that different judges would make a different decision for the same individual — but we have no evidence on the relative importance of inconsistency versus cross-judge variation in leniency in generating unfairness, and no methods to estimate the prevalence of inconsistency. In this

paper, I show how inconsistency can be identified at the judge level by exploiting multi-stage decision processes and using the judges to check the quality of each others decisions.

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

I implement the model using data on judicial review of initially-denied refugee claims at the Federal Court of Canada. Although the justices of the Federal Court are experts in refugee cases, I uncover relatively high levels of inconsistency. For first-round judges who approve the same share of cases (on average, 14%), I estimate they disagree on 13.2% of cases, and bound disagreement to at least 3.6% of cases. Even more strikingly, inconsistency contributes more to unfairness than cross-judge variation in leniency. If all judges were perfectly consistent they would disagree on 8% of cases, but with the addition of inconsistency the disagreement rate jumps to 23%. Overidentification tests suggest that most disagreement arises from idiosyncratic observational errors, rather than permanent differences between judges in which aspects of cases they think are most relevant.

Cross-judge variation in inconsistency is large. I validate the measured variation against a survey that solicited estimates of judicial characteristics from refugee lawyers who had appeared in front of the judges. Judicial consistency improves dramatically after the first year, and

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

It is unclear whether the overall level of inconsistency I uncover here would be found in other courts. By construction the Federal Court's caseload is difficult, consisting of initially-denied refugee claimants who appeal the decision, and the lack of precedent likely increases inconsistency. That said, the improvements in consistency over time, and with respect to workload and the judge selection mechanisms are much more likely to transfer to other contexts.

In the Appendix, I show that this level of inconsistency implies relatively large levels of bias in MTE estimates using judge-assignment instruments, but not for linear IV estimates. Future work should determine whether the results hold in criminal courts and other decision-making institutions such as the Social Security Administration, and once the econometrician can condition on type of case and other covariates. If the level of inconsistency at the Federal Court is also present in other contexts, it would introduce bias into estimates of the effect of incarceration (Aizer and Doyle, 2013; Mueller-Smith, 2014), SSDI receipt (Maestas et al., 2012), and patent receipt (Gaulé, 2015) recovered from examiner-assignment IV designs.

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

CHAPTER 3

The Effects of Parental and Sibling Incarceration: Evidence from Ohio

3.1. Introduction

The United States has the highest rate of incarceration in the developed world, directly affecting millions of prisoners annually. Even more individuals are indirectly impacted by the criminal justice system, as friends, co-workers and family members of the incarcerated. The impact is particularly large for children in marginalized groups – around 25% of black children and 30% of children of non-college educated parents experience the incarceration of a parent by the age of 14 (**?**). Given the large number of people affected by these spillovers, the indirect effects of incarceration could potentially be even larger than the direct effects.

While the impacts of sibling and parental incarceration are theoretically ambiguous, the prevailing wisdom has been that it will have negative repercussions (**?**). Advocates and academics alike argue that the trauma associated with familial incarceration, combined with the removal of social and economic support, will cause children to engage in a number of harmful behaviors, such as disengagement from school and crime. On the other hand, there are several plausible mechanisms through which the indirect effect of incarceration could be positive. Parents who are incarcerated may be lower quality caregivers than their replacements, or in cases of abuse, commit crimes that directly and adversely affect their children. Similarly, siblings may commit crimes together or a sibling who engages in criminal activities may influence the others to emulate this behavior. More generally, the intrafamily effects of incarceration are likely to

vary depending on the relationship between the family members, and the net indirect effect of incarceration is unclear.

Despite the importance of this topic, empirical evidence on the relationship between incarceration of family members and child outcomes in the United States has been largely correlational (**?**). The key empirical challenge is that the children of incarcerated parents come from systematically more disadvantaged home environments than the children of non-incarcerated parents. To the extent that unobserved dimensions of home environment are correlated with parental incarceration, observational studies are biased towards finding a negative effect of parental incarceration. Perhaps because of these methodological issues, most studies find negative effects in the short-run on outcomes such as antisocial behavior (**?**), drug use (**?**), academic achievement (**?**), and criminality (**?**). A few studies use panel data to estimate the effect of parental incarceration on child academic achievement in a diff-in-diff framework and suggest minimal (**?**) or even positive effects (**??**) of parental incarceration.

The lack of causal evidence is largely due to the stringent data requirements. To generate non-correlational estimates, it is necessary to have criminal justice data with a source of exogenous variation in incarceration, data that links family members, and outcome data for the family members. This data must span a period of twenty to thirty years in order to measure adult outcomes for a sufficiently large sample of family members. This paper overcomes these challenges and provides the first causal estimates of the spillover effects from incarceration on family members in the United States. We collect and merge adminstrative data sets from 14 separate government departments for the three largest counties in Ohio, containing a population of approximately 3.4 million people: Cuyahoga County, which contains the city of Cleveland, Hamilton County, which contains the city of Cincinnati, and Franklin County, which contains

the city of Columbus.[1] We reconstruct families by linking defendants to their children using birth certificates, and to siblings by matching through their common parents.

In Ohio, cases are randomly assigned to judges who differ in their propensity to incarcerate defendants. We leverage random assignment as a powerful source of exogenous variation in incarceration probability. In our sample, a lenient judge at the 10^{th} percentile of strictness incarcerates only 23% of defendants, whereas a judge at the 90^{th} percentile of strictness incarcerates 52%. For each defendant, we instrument for incarceration with their judge's average propensity to incarcerate *other* defendants, following a strategy used in a number of recent papers (**?????**). This instrument is strong and uncorrelated with the characteristics of cases or defendants assigned to the judge.

We generate causal estimates of three treatment effects: (1) the direct effect of incarceration on defendants, in order to contextualize the ways family members are affected; (2) the effect of parental incarceration on children; and (3) the effect of incarceration of siblings. For defendants, incarceration leads to large short-run decreases in criminal activity during the time spent behind bars, but after release, prisoners commit crimes at a higher rate than non-incarcerated defendants. After five years, the cumulative number of new crimes committed by the two groups is similar.

Given that we do not find rehabilitative effects on those who are incarcerated, the literature would suggest that parental incarceration will harm children by separating them from their primary caregivers. Instead, we find that the impact is largely positive. Incarceration of a parent

[1] Analysis of some of the data is still in progress and will be included in a May 2018 draft of the paper. In addition to more outcomes from Ohio, that draft will include data from a fourth location: Allegheny County, Pennsylvania, which contains the city of Pittsburgh. The Allegheny county data is particularly rich, and includes outcomes such as mental health, substance abuse, and child welfare services involvement. Data usage agreements are also in progress for use of unemployment insurance data to measure employment outcomes of formerly incarcerated individuals and their family members.

reduces the child's probability of being incarcerated as an adult by 3.1 percentage points (37 percent of the mean adult incarceration rate) and as a juvenile by 2.9 percentage points (63 percent of the mean juvenile incarceration rate). The effects are even larger when focusing on children of incarcerated mothers, who are 7.2 percentage points less likely to be incarcerated as adults and 6.4 percentage points less likely to be incarcerated as juveniles.[2] Results are qualitatively similar across male and female children, though are generally larger for male children.

We examine several other policy-relevant outcomes — teenage pregnancy and child educational attainment — for potential negative effects of parental incarceration. With a high degree of precision, we can rule out increases in becoming a parent as a teenager for both boys and girls, including among those with the largest reductions in criminal activity, children with incarcerated mothers. We similarly do not detect any effects on school attendance or child standardized test scores in reading or math, though the estimates are less precise. Given the lack of negative results and large reductions in criminal justice system involvement, it appears that incarceration of parents actually improves child outcomes.

Spillover effects are similar for siblings, where the incarceration of a sibling with a shared mother leads to a 7.3 percentage point decrease in the probability of a child being incarcerated as an adult. Incarceration of a sibling may have a positive effect if siblings are likely to engage in criminal activities together (**?**). More generally, a sibling engaged in criminal activity may influence other children in their family towards criminal activity, an effect that their removal will attenuate.

[2]The point estimates are even larger when looking at the outcome of whether the child appeared in court as an adult or juvenile. However, we prefer to focus on the outcome of incarceration since this suggests that the individual actually committed the offense for which they were charged.

From a policy perspective, it is helpful to understand how different subpopulations react to the incarceration of family members. It may be that effects are worse among poorer or more vulnerable children, as found in **?** in Sweden. Using geocoded address data for the children, we find the opposite, where the lower the socioeconomic status of the neighborhood of the child (at the census block group level), the larger the reduction in future criminal justice system involvement as the result of the incarceration of a family member. Effects are particularly large in the case of siblings, where for children in the poorest quartile, the incarceration of a sibling with a shared mother reduces own probability of incarceration by 21.7 percentage points.

There are a number of mechanisms that could plausibly explain our results. First, one might imagine that parents are rehabilitated by their experience of incarceration, but that is inconsistent with the increased levels of criminality we observe post-release. Another explanation is that children are "scared straight" by the experience of seeing a family member being incarcerated. That would be consistent with the reduction in future criminal justice system involvement, and lack of change on other margins of risk behavior, such as teen pregnancy. Finally, many parents on the margin of incarceration may be of a lower quality, and so the child may be better being removed into other, possibly more stable home environments (e.g. with grandparents).

The effect of familial incarceration may vary along dimensions that can be manipulated by policymakers, such as sentence length. For each charge we observe in the data (e.g., assault), we calculate the average sentence length conditional on incarceration, and measure how the effects of incarceration vary based this measure of potential length of sentence. The positive effects of incarceration are largest among children whose parents have shorter expected sentences, while for children whose parents have expected sentences of more than a year, there is no effect of parental incarceration on child crime outcomes. This is suggestive evidence of

a "scared straight" effect that is triggered even for short sentence lengths. If the results were caused by placing children with higher quality care-givers, one would expect that longer sentences would lead to even larger positive effects, rather than smaller ones. This test is not entirely satisfactory, since children of parents with longer sentences may also tend to be systematically disadvantaged in other ways, and so the relevant heterogeneity may be along that dimension. Future drafts of the paper will provide stricter tests of the mechanisms at work, including estimating heterogenous effects along estimated parent quality and using child welfare data to determine the caregivers of children after their parents are incarcerated.

The first literature to which this paper contributes is the study of the indirect effect of incarceration on child outcomes. Outside of economics, this has been an active area of research for decades, but all research has used correlational research designs such as multivariate regression (**?????**) or matching to construct plausible "control" groups of children whose family members were not incarcerated (**?????**). These strategies do not control for unobserved heterogeneity in home environment that is related correlated with incarceration, and are unlikely to generate reliable estimates. There has been a recent flurry of interest among economists on the effects of parental incarceration, taking advantage of large administrative datasets and focusing on the incarceration of fathers. **?** compares children during as well as before and after parental incarceration and concludes that incarceration improves academic achievement and reduces behavioral issues. Two other recent papers use administrative data from Scandinavia and follow a similar empirical strategy as our paper, exploiting random assignment of judges to cases to generate exogenous variation in the probability of incarceration (**??**). These papers find highly negative and null effects of incarceration on child outcomes, respectively.

In contrast to previous work, this is the first paper to provide quasi-experimental results from the US context, as opposed to Scandinavian countries (**??**). Estimates from the United States have broader policy relevance since there are currently over 2 million incarcerated individuals in US jails and prisons, as opposed to only 6,000 in Sweden and 4,000 in Norway (**?**). In Ohio alone, there are currently 70,000 individuals in either jail or prison despite having a population only slightly larger than that of Sweden (**?**). Scandinavian justice systems are significantly more rehabilitative than those in the United States, and so estimates from those contexts may be hard to apply to the US. For example, spending on inmates in Swedish and Norwegian prisons averages over $120,000 per year, versus $30,000 in US prisons. **?** find that Norwegian prisons have an economically rehabilitative effect on those incarcerated, as opposed to **?**, who shows that incarceration in the US has sharply negative economic effects.

Second, the size of our data allows us to broaden the scope of the study. As a result of high incarceration rates in the United States and the long time frame of our data, we observe approximately one million unique defendants. This allows us to study the effect of incarcerated siblings, as well as generate precise estimates of effects of familial incarceration for various policy-relevant groups, such as socio-economic status. To date, little attention has been paid to the incarceration of siblings despite the police relevance of this phenomenon. A representative survey of US inmates found that 19% had a father who was ever incarcerated, 6.4% had a mother who had been incarcerated, but a massive 34.4% had a brother who had been incarcerated (**?**). Furthermore, given the likely heterogenous effects of incarceration of family members, understanding how the effects may vary is vital.

Third, our findings are in sharp contrast to nearly all previous work. For example, **?**, who find large negative effects of incarceration in Sweden, speculate that the negative consequences

of parental incarceration for children must be even larger in the United States given the punitive nature of the American justice system. Our estimates have important implications for estimating the costs of incarceration and considering the likely effect of sentencing reform. We find that short sentences can actually have positive effects on family members of incarcerated individuals. From the perspective of family members of the incarcerated, a shift towards shorter sentences may be a more beneficial path for sentencing reform than reducing the frequency of incarceration.

The second literature we contribute to is on the determinants of youth criminal and delinquent behavior. Previous work has found many significant determinants of these activities, such as peer influences (**????**), assignment to foster care (**??**), prior stints of juvenile incarceration (**?**), and education (**??**). We add to this literature by showing another channel through which the justice system indirectly affects criminal behavior.

Finally, we contribute to the large literature on the effect of family on children's economic outcomes. This includes an enormous literature on how parents affect their children (**???**), shocks to parents (**???**), shocks to siblings (**?**), and order of birth (**??**). Given that children are being separated from their parents as a result of incarceration, our results stand in contrast with those of **?** and **?** on foster care. Those papers find removal from immediate family has large negative effects on adult outcomes, while we find that incarceration shocks *positively* affect family members. Parents who are incarcerated may be worse caregivers than those whose children are removed by foster care or familial support may be greater for children post incarceration incidents (e.g. placement with grandparents).

Section 3.2 provides a more detailed discussion of the theoretical spillover effects of incarceration, as well as background on the courts systems in Ohio. Section 3.3 describes our

data, Section 3.4 describes the empirical strategy, and Section 3.5 shows the results. Section 3.6 concludes.

3.2. Background

The US incarceration rate is five to ten times higher than in most other industrialized democracies (**?**). This is a relatively new phenomenon; the state and federal prison population rose from about 200,000 in the early 1970s to 1.5 million in 2009, with an additional 700,000 held daily in local jails (**?**). Much of this has been attributed to longer sentencing and mandatory minimum sentencing, which grew in popularity in the 1980s (**?**). As a result, more households than ever experience the incarceration of a family member, and for longer periods of time. Poor communities and African-Americans have been disproportionately affected by these changes, with African Americans are incarcerated at more than five times the rate of whites (**?**).

Higher rates of incarceration may have lowered crime through incapacitating those who are likely to commit crimes or deterring would-be offenders, but may also lead to more crimes being committed by the released prisoners, who gain crime-specific human capital while in prison. The largest, most rigorous study in the US context concludes that incarceration increases overall crime rates (**?**). However, focusing solely on current crime rates ignores the potential spillovers on peers or future generations. Between 1980 and and 2000, the number of children with an incarcerated father rose from 350,000 to 2.1 million, around 3% of all US children (**?**).

There is an extensive and growing literature in sociology and criminology that examines the collateral consequences of the incarceration of family members. The vast majority of this work has focused on the effects of incarceration of parents, since their incarceration is likely to be particularly impactful for a child. **?** lay out three theoretical channels through which the effects

may be harmful: strain, socialization and stigma. The strain of the loss of an economic provider may lead to damage in the short and long run. Incarceration reduces household income, which may extend after the prison spell has ended, due to lower human capital and social stigma. **?** finds that each additional year in which an adult is incarcerated reduces employment after release by 3.6 percentage points. The effects are even larger for those with stable earnings prior to incaceration. This loss in income may have negative effects on education and other human capital investments — in a recent study on the effects of cash grants for poor to middle-income families, a $1,000 increase in 1995 family income increased math and reading tests scores by 0.04 SD (**?**).

The loss of a source of social support (socialization) may impede child development. A long line of research in psychology on attachment and social bonding theory points to separation as damaging to children (**??**). A detailed, small-scale study of the effects of maternal incarceration confirms that initial separation with the mother generates widespread negative emotions (**?**). Other research suggests that the reduction in parenting inputs may extend past the duration of incaceration: a study by **?** suggests that incarceration is associated with marital dissolution and **?** find the incarceration of a partner harms the mental health of a child's other caregivers. The stigma to the child of having an incarcerated family member may lead to social isolation that catalyzes criminal activity on the part of the child. Incarceration of family members may also lead children to follow in their footsteps of criminality.

Conversely, incarcerated people may be lower-quality caregivers than their replacements. In this case, incarceration could increase certain dimensions of caregiving inputs and improve long-term outcomes. Family members who are offenders may sap economic and social resources and prove a disruptive rather than supportive influence, inhibiting the development of

pro-social attitudes in a child (e.g. **?**). In some cases, such as for those accused of physical or sexual abuse, the crime committed has a direct negative effect on children. Children's exposure to the incarceration of a family member may also cause them to become "scared straight", i.e. perceive the expected punishment for criminality to be higherand decrease their level of subsequent criminal activity (**?**). If family members (and particularly siblings) are committing crimes together, then the incarceration of one member may reduce the criminal involvement of the other members.

On net, empirical research has typically found that the long-term effects of the American system of incarceration are negative, although other studies have found positive effects in less punitive Scandinavian systems (**??**). Parental incarceration has been shown to have negative effects on outcomes such as the child's criminal justice system involvement (**??**), mental health and learning disabilities (**???**), academic achievement (**?**), substance abuse (**?**), and physical health and obesity for female children (**?**). Others have found no effect of parental incarceration on behavioral issues (**?**), substance abuse (**?**) or test scores (**?**). In most of the quantitative studies, the results are from regressions across a population of children where some have incarcerated parents and others do not. This shows whether there is a correlation between the incarceration of the parent and outcomes for the child. In the highest quality studies, control variables are included in order to attempt to account for pre-existing differences between children with incarcerated parents and those that are not incarcerated. For example, children with parents who are incarcerated will tend to have lower family income even prior to incarceration. Thus the study might control for pre-existing family income, although it is impossible to control for all confounding factors (**?**).

Recent work has challenged the paradigm of conceptualizing incarceration of family members as a net "good" or "bad". For example, **?** argues that in some cases, it will remove a harmful influence, while in others, it takes away a valuable form of social support. Other recent studies have found that whether the effect of incarceration is positive or negative depends on factors such as the child's environment, the nature of the relationship between the prisoner and child, and the gender of the parent (**???**). In some cases, the incarceration of a parent may have no effect on a child, such as if the ties between the child and parent are weak (e.g. an absentee father). In others, such as the removal of a primary caregiver (especially mother), the effect may be quite negative. Aside from providing more credible causal estimates, one of the key open questions in the literature is identifying the dimensions of heterogeneity that mitigate or exacerbate the collateral consequences of incarceration.

3.2.1. The Criminal Justice System in Cuyahoga, Franklin, and Hamilton Counties

This study investigates incarceration in the context of the three largest counties in Ohio: Franklin County (population of 1.3 million), Cuyahoga County (population of 1.2 million), and Hamilton County (population of .8 million). These counties each contain a sizeable urban core surrounded by outlying suburbs, and have similar median household incomes of around $41,000 per year. The counties are also demographically similar, with populations that are approximately 70% white and 25% black.

Ohio is a particularly relevant state in which to study the intrafamily spillovers of incarceration, since it is has one of the highest rates of parental incarceration in the US. One in ten Ohio children have a parent who has been incarcerated; only two others states have higher rates (Kentucky and Indiana) (**?**). These counties are also broadly representative of crime and

incarceration policy in the United States. Among the 84 US cities with populations of at least 250,000, Cincinnati ranks 10th, Cleveland 18th, and Columbus 37th in standardized violent crime rates as measured in the FBI Uniform Crime Reporting Statistics (2014).[3]

In each county, the justice system is divided into Municipal and Common Pleas Courts. Municipal courts are responsible for misdemeanor criminal and traffic cases, and each county has 20,000-30,000 municipal cases per year. The most common types of offenses that come before Municipal Courts are misdemeanor drug possession (13.3% of cases), misdemeanor theft (8.5% of cases) and disorderly conduct (7.5% of cases). As a function of the less serious nature of these crimes, incarceration is relatively rare, and only 14.9% of defendants are immediately sentenced to incarceration.

Felony cases are decided in the Common Pleas courts.[4] Each county handles between 5,000 to 15,000 cases per year, of which 26% are serious drug offenses such as trafficking or possession of cocaine or heroin, and 18% are made up of felony theft, burglary and robbery. In 34.4% of cases, the defendant is sentenced to incarceration, and in 36.4% sentenced to some sort of alternative to incarceration, such as probation.

In both Municipal and Common Pleas courts, a single judge is responsible for managing all aspects of the case, including negotiations over plea agreements and sentencing. This judge is assigned as soon as possible after arraignment, and by Ohio law, assignment must be random. This rule was put in place to avoid judge-shopping, in which individuals manipulate the assignment system to receive a judge who might be more favorable to their case. The rule specifically mandates true random assignment and specifies a list of acceptable and random system, such as

[3]Standardized violent crime measures include murder and non-negligent manslaughter. Among all violent crimes, the rankings are similar: Cleveland (9th), Cincinnati (26th), and Columbus (55th).

[4]The Common Pleas courts are also responsible for domestic relations, juvenile and probate courts, but these records are not relevant for this paper.

a drawing balls from a bingo cage.[5] Assignment is currently carried out by a computer system, which we will later validate as random by showing that the characteristics of cases are unrelated to the judge to which they are assigned.

Both Municipal and Common Pleas judges are elected on a non-partisan ballot for six-year terms. All candidates must be attorneys with at least six years of legal experience, and elections are typically quite competitive. Restricting our sample to judge who hear at least 100 cases between 1990 to 2017, we observe 212 unique Common Pleas and 107 unique municipal judges. This works out to approximately 1,009 cases per year for municipal judges and 388 cases for common pleas judges, reflecting the more complex nature of common pleas cases. Over the sample period, the average municipal court judge in our sample hears 9,396 cases, while the average common pleas judge oversees 3,735 cases.

3.3. Data

We combine administrative data from a variety of sources. In Ohio, adult court cases are a matter of public record, and all court records are available on the websites of the respective County Clerks of the Court. For each county, we scraped all Common Pleas Cases between 1990 and 2017, a total of 765,000 records. The current draft includes the municipal data for Franklin and Hamilton Counties over the same time period (1.14 million records), and future drafts will also include data for Cuyahoga county. This is the full sample of cases, and so includes cases that were dismissed or in which the defendant was determined to be not guilty.[6] The case records contain the full docket history of all transactions that occurred on the case, including the filing

[5] Aside from capital cases, which are omitted from our analysis, it prohibits the use of any quasi-random systems that could be manipulated, such as assigning cases by rotation to judges (i.e. judge 1 gets the first case, judge 2 gets the second case, and so on).

[6] The only cases that are not present are those in which defendants have filed for expungement. These have been removed from all databases, but are less than 1% of cases.

of charges, assignment of judge, and sentencing. They also include defendant characteristics such as name, date of birth, gender, race and home address.

In order to match defendants to their family members, we use birth certificate records from all births in Ohio between 1970 and 2017 (approximately 5 million records). The birth certificates contain the full name and date of birth of the child, as well as the name and additional information on both the mother and father. The mother data is available for virtually all births; information about the father is missing on 12% of records.[7] The data includes residential address at birth, which we match to 2011-2016 American Community Survey information at the census block group level, generating information about the socioeconomic status of the neighborhood. This provides a measure of how family involvement in the criminal justice system may interact with the socioeconomics status of study individuals and their communities. Furthermore, the work of **?** and others show that neighborhood of residence is itself an important outcome given that neighborhood characteristics are influential for child outcomes.

The court records contain the full name and date of birth of the defendants, so we begin by using this information to match defendants to the parents in the birth records, allowing for inexact matches on name. With the parent-child connection in hand, we match the children forward 1) to court records to measure criminality; 2) to birth records to measure teen pregnancy; 3) to school records to measure academic achievement, and 4) to voter records to measure civic engagement and Ohio residency. Similarly for siblings, we match defendants to their own birth record, then find all other children with the same parents. Since siblings may not share the same set of parents, we differentiate between full and half siblings, and whether the common parent is the mother or father. We describe the matching process in detail in the Online Appendix.

[7]This missing data is not terribly problematic since fathers who are missing on the birth certificate are those least likely to be involved in their childrens lives.

The schools data as well as data on juvenile courts were only available in Cuyahoga County, and we signed data sharing agreements with the Cleveland Metroplitan School District and the Cuyahoga County Juvenile Courts. The schools data contain student test scores, attendance, grade completion, and graduation for all students between 2010-2017. The juvenile courts data contain all cases between 1995 and 2017, allowing us to measure whether a child was involved with the juvenile justice system and the seriousness of their offense.

The voter records represent the universe of Ohio voters between 2000 and 2017. While voter registration and actually voting in elections are a key measure of civic engagement and, hence, important in their own right, we primarily use these data to test whether the study sample has stayed in Ohio and their within-state migration patterns by geocoding the addresses.

We have additionally collected data from Allegheny County, Pennsylvania, which contains the city of Pittsburgh. This is the second most populous county in the state of Pennsylvania (population of 1.1 million), only slightly smaller than the county of Philadelphia (population of 1.3 million). We have merged data from eight unique administrative data sets[8]. We follow the same empirical strategy as in the three Ohio counties, using birth certificates to match parents and children and taking advantage of the random assignment of cases to judges for variation in likelihood of incarceration. Analysis on this data set is in progress and will be included in the next draft of the paper. Pending completion of data usage agreements, future drafts will also include data on employment and earnings for incarcerated parents and their children.

[8]These are: 1) Birth certificates for all births between 1975 and 2017 in Allegheny County; 2) Court records for all cases between 1977 and 2017; 3) Juvenile court records between 2007 and 2017; 4) School records from Pittsburgh Public Schools between 2004 and 2017, including attendance, grades, test scores, and disciplinary incidents; 5) Records of all child welfare cases handled by the Allegheny County Department of Human Services between 2002 and 2017, including allegations of abuse against parents and removal; 6) Probation and parole records between 2008-2017; 7) Allegheny County Jail records between 2002-2017; and 8) Department of Human Services records of substance abuse and mental health issues

3.3.1. Descriptive Statistics

Table 4.13 displays the characteristics of cases that come before the Common Pleas courts in each county. Although the counties are all predominantely white, a majority of defendants in each county are black. At the time that charges are filed, half of defendants are below the age of 30, with 25% below the age of 23 and 25% above the age of 40. Defendants are disproportionately male (77% of cases). In all of the counties, property crimes and drug crimes are the most common types of offenses. Although 48% of cases feature first-time defendants, many defendants have multiple previous offenses. Incarceration rates also vary greatly between the counties, with higher rates in Hamilton and lower rates in Cuyahoga. Conditional on being incarcerated, the average sentence lengths are fairly similar across the counties, though those in Hamilton are slightly longer.

Using address information on the court records, we find that the average defendant comes from a neighborhood in which 39.6% of households are below the poverty line and 31% of households are beneficiaries of the Supplemental Nutrition Assistance Program. To visualize the extent to which incarceration disproportionately affects poor households, we calculate the fraction of residents below the poverty line in each census block group in Ohio and rank them from poorest to wealthiest. Figure 4.23 displays a histogram of the distribution. If defendants were randomly drawn from the population, the distribution would be uniform, but unsurprisingly incarcerated individuals are strongly concentrated in poorer neighborhoods. Half come from the poorest 16.9% of neighborhoods. [9]

[9]This figure is similar when looking at the total number of defendants, where half of defendants come from the poorest 19.4% of neighborhoods

After matching the court records to birth certificate data on parents, Table 4.14 compares the sample of defendants who are parents to those that are not. Most of the differences between the populations are small and not economically meaningful. Parents are two years younger, and accused of slightly less serious crimes as measured by the mean sentence for the charges that have been filed against them.

3.4. Empirical Strategy

Our goal is to estimate the effect of parental and sibling incarceration on a range of child outcomes. A naive approach would be to regress child outcomes on whether their parent were incarcerated. This strategy is unlikely to generate causal estimates because incarceratoin is not randomly assigned. A child whose parent has been incarcerated for 5 years for aggravated assault is likely to differ in many observable and unobservable ways from a child whose parent has been charged and acquitted of marijuana possession. We instead instrument for whether the parent is incarcerated with the identity of the judge. As discussed in Section 3.2, Ohio state law mandates that judges are randomly assigned to cases. If this is indeed true, and if judges differ in their propensity to incarcerate, then judge identity is a valid instrument for incarceration. A standard IV regression model for this problem takes the following form:

$$y_{ijct} = X_{ijc}\beta + \phi I_{ijct} + \gamma_{ct} + \varepsilon_{ijct} \tag{3.1}$$

$$I_{ijct} = X_{ijc}\alpha + \lambda z_{(i)j} + \mu_{ct} + e_{ijct} \tag{3.2}$$

for individual i who has been assigned to judge j of county c in year t, where y_{ijct} is the outcome of interest, X_{ijc} is a vector of controls, γ_{ct} is a county-year fixed effect, I_{ijct} is the endogenous decision to incarcerate the defendant, and $z_{(i)j}$, is an instrument for that decision. To be a valid instrument, $z_{(i)j}$ must be related to the endogenous variable of interest I_{ijct}, but unrelated to confounding factors ε_{ijct}.

As is common in the judge-effects literature (e.g., **?**), we instrument for incarceration with information about how the judge has treated other defendants. Specifically, we take the mean incarceration rate for all other cases, after residualizing out observable case characteristics to increase precision. Averaging over all other cases removes the mechanical correlation between own outcome and judge-average outcome for judges with few cases.

$$z_{(i)j} = \frac{\sum_{k=1}^{N_j} \mathbb{1}[k \neq i](I_{kj} - X_{ijc}\widehat{\pi} - u_{ct})}{N_j - 1} \tag{3.3}$$

Note that since parents are only a fraction of the overall sample, we calculate this leave-out mean over the entire sample of defendants rather than the sample of parents. This increases precision of the instrument and reduces possible finite sample bias, since there are more cases on which to estimate the leave-out mean for each judge.

3.4.1. First Stage

Figure 4.24 presents a histogram of the instrument. The value of the judge instrument varies from approximately -0.12 to 0.18, and for each 0.1 increase in the instrument, the likelihood of incarceration increases by approximately 9 percentage points. The coefficient of the instrument on defendant incarceration is less than 1 due to measurement error - we observe a finite number

of cases per judge, so the first stage is biased towards zero. The first column of Table 4.15 presents the linear first stage of Equation 3.2. Robust standard errors are clustered at the judge level. As suggested by Figure 4.24, the instrument is strong, with an F-statistic of about 2000.

Instrumental variable regressions identify the local average treatment effect (LATE), the average treatment effect of "compliers" (**?**). In this case, that is the sample of children whose parents are incarcerated on the basis of being assigned to a stricter judge. It is not possible to identify specific individuals in the data who are compliers; however, it is possible to describe some of their observable characteristics by re-estimating the first stage relationship across multiple subsamples (**?**). If the instrument has a stronger (weaker) relationship with incarceration in a particular subsample, this implies that compliers are more (less) heavily concentrated in that group. If the instrument Z is binary, endogeneous treatment is I, and binary covariate is X, then (**?**):

$$\frac{P[X=1|complier]}{P[X=1]} = \frac{P[I|Z=1,X=1]-P[I|Z=0,X=1]}{P[I|Z=1]-P[I|Z=0]} \tag{3.4}$$

In other words, the ratio of the complier share of the demographic group to the overall share of the demographic group is equal to the relative first stages of the demographic group and the overall population. The same intuition holds with our continuous instrument. For the remaining columns of Table 4.15, we estimate the first stage by subgroup and calculate the ratio of the first stages. For most subgroups, this ratio is indistinguishable from 1, suggesting that this subgroup makes up an equal share of the complier population as the overall population. One notable exception is individuals accused of low-severity crimes (we divide charges into tercile of mean sentence conditional on incarceration), who are less likely to be compliers, and individuals who

are accused of medium-severity crimes, who are more likely to be compliers. The coefficient for parents is almost identical to the full-sample estimate, suggesting that parents are no more or less likely to be compliers than the general population. An additional benefit of running these tests is that finding a positive relationship between the instrument and the endogenous regressor across various subsamples of the data provides support for the monotonicity assumption.

3.4.2. Judge Assignment and Instrument Validity

Leave-out judge mean strictness must satisfy the exclusion restriction to be a valid instrument. The main concern in this context is that cases might be assigned to judges non-randomly, such as if one judge were known to be particularly strict, and well-informed defendants were able to manipulate the justice system to avoid being assigned to them. In that case, the leave-out mean for the judge would be a combination of their underlying strictness and the fact that the cases being assigned to them were for systematically less well-informed defendants. Estimates of the effect of judge strictness (and through it, incarceration) will be biased. For example, if less well-informed defendants tend to be poorer, it will appear that the stricter judge is causing them to be poor, when in fact poorer defendants are simply more likely to be assigned to a strict judge.

As described earlier, this concern is mitigated in Ohio since state law mandates that judges be assigned randomly to cases. Because of random assignment, the characteristics of defendants assigned to a judge should be uncorrelated with the underlying strictness of that judge. However, there are a few exceptions that must first be taken into account. Capital cases are non-randomly assigned due to their greater sensitivity and requirement of resources. There are also specialty courts that focus on certain types of cases, such as those in which the defendant is a military

veteran. These specialty courts typically only have one judge who oversees all cases falling under the court's jurisdiction. Most importantly, if a defendant has an active case in front of a judge, or if they have been previously sentenced and are on probation, new cases are assigned to the original judge. We drop all of these non-randomly assigned cases from our sample.

Cases are also sometimes transferred between judges. This occurs most often upon the retirement of a judge, after which most of their cases are transferred to their replacement. New judges by definition have lower levels of experience and may systematically differ in their sentencing tendencies. To be conservative, we consider the judge assigned to a case to be the original judge who was assigned to the case, at the time of random assignment. This weakens the strength of the instrument, since the first judge is no longer the sentencing judge on the case, but appropriately matches our empirical strategy to the source of randomization.

One implication of randomization is that defendant characteristics should be uncorrelated with judge severity. We test this implication in Table 4.16. Columns (1-3) show that the types of the cases assigned to the judge are unrelated to the judge's strictness. Stricter judges are no more or less likely to be assigned particular types of cases, such as drug cases (column 1). Column (2) takes the most serious charge in a case and calculates the leave-out average sentence for that type of charge in the court, in order to measure to seriousness of the case. Since there is a long right tail of sentence lengths, column (3) takes the log of the leave-out average sentence. Neither is related to judge strictness. Columns (4-6) show that the judge strictness is also unrelated to either the gender, race, or age of the defendants assigned to them. The seventh and eighth columns use American Community Survey data on the neighborhood from which the defendants come to proxy for their level of income. Neither the neighborhood median income nor percentage of neighborhood members who receive SNAP are related to the

strictness of the judge. All of the evidence points to the randomization process having been properly adhered to, and thus that the judge assignment is exogenous.

3.5. Results

3.5.1. Direct Effects of Incarceration

In order to understand the indirect effects of incarceration, it is first necessary to understand the direct effects. In theory, time spent in a pentitentiary could strengthen fear of punishment, teach positive skills such as literacy or job skills, or help individuals break addiction to drugs. **?** find exactly this effect in Norwegian prisons, but this system has a substantially greater focus on rehabilitation than in the United States. A few studies in the United States have also found that incarceration can reduce future criminality (**?**), but the majority of high-quality studies find that time spent in prison increases future criminal behavior (**????**) through channels such as building criminal capital and networks (**?**). **?** shows that incarceration also substantially worsens economic outcomes for the incarcerated individual over the long-run.

Figure 4.25 shows how incarceration affects future criminal behavior in our sample. Panel A plots the coefficients from a regression of incarceration on the judge instrument for each of the 30 quarters following judge assignment. Note that incarceration peaks in the second quarter, reflecting the time it takes for cases to wind their way through court. Panel B displays coefficients from a similal quarter-by-quarter regression of cumulative number of new charges on judge instrument. Corresponding with the period of incarceration, there is a short-run dip in additional criminal charges from the incapacitation of incarcerated defendants. After approximately 7 quarters, the trend reverses and reverts to zero or possibly additional charges. This suggests a period of initial incapacitation of incarcerated individuals, reducing crime, but after

they are released, they converge to the level of crimes committed by those who were not incarcerated. This is consistent with a criminogenic effect of prison, and implies that incarceration only delays criminal activity into the future for these defendants.

3.5.2. Spillovers of incarceration on children

The vast majority of papers in the literature suggest that parental incarceration is harmful to children. One of the two causally-identified studies to date, **?** finds that even in rehabilitative, Scandinavian style prisons, children experience a great deal of harm from having an incarcerated parent. The second, **?**, does not find effects that are significantly different from zero, but suffers from large standard errors. Both papers state that in the United States, where the system of justice is more punitive, the effects of parental incarceration are likely to be more negative than Nordic counties. Indeed, a non-causal, cross-sectional regression finds that children with incarcerated parents are signficantly more likely to be charged with a crime and are more likely to be incarcerated as adults (**??**).

Table 4.17, showing the causal effect of parental incarceration, finds the opposite. Panels A and B test whether having an incarcerated parent makes a child more or less likely to appear in court or be incarcerated as an adult. For both outcomes, the effects are marginally statistically significant in the full sample, but strongly significant when broken down by gender of the incarcerated parent. Children with an incarcerated mother are 10.8 percentage points less likely to be charged with a crime as an adult, and 7.2 percentage points less likely to be incarcerated. The effects appear stronger for male than female children, although the differences between the genders are not statistically significant.

The effect of having an incarcerated father is weaker, but more consistent with conventional wisdom. Girls with an incarcerated father are 6.1 percentage points more likely to appear in court as an adult and 3 percentage points more likely to be incarcerated themselves. The weaker relationship between paternal incarceration and child outcomes is consistent with fathers being less likely to be involved in the lives of their children. In future drafts of the paper, we will take parent addresses from court records to check if they are coresident with their children, based on the child's address in school records. It may be that the relationships are stronger for parents, and fathers in particular, who live with their children.

Panel C does not find any statistically significant relationship between parental incarceration and whether the child becomes a parent as a teenager. This is determined by whether the child appears as a parent on birth certificates prior to their 18^{th} birthday. The standard errors in the main specification are fairly small, and exclude effects of policy relevant sizes.

Why does having an incarcerated parent reduce a child's future involvement with the criminal justice system? It may depend on the nature of the incarceration episode. Short stints of incarceration could "scare the child straight" by making incarceration a more salient result of crime. On the other hand, longer stretches of incarceration may lead to such large reductions in parental inputs that the overall effect is negative. As a test of this theory, we calculate the expected sentence length conditional on incarceration for each charge (truncating at age 18 for each child), and test whether the effect of incarceration differs along this dimension. Table 4.18 tests this by linearly interacting potential exposure to parental incarceration with incarceration (parametric versions of this chart are available in the Online Appendix. See **??**, **??**, and **??**).

Broadly, Table 4.18 shows that short stints of incareration, below the 50th percentile of exposure, reduce a child's likelihood of future criminal justice system involvement. The effect

sizes are sizeable, and concentrated more among male children than female children. However, exposure to longer sentences reverses the effect. Among children whose parents are incarcerated with sentence lengths averaging above the 90th percentile of length, children are more likely to become teen parents. The effect of teen pregnancy is concentrated among girls, and nearly doubles the probability of teen pregnancy for the most exposed female children.

A potential issue with this approach is that the exposure measure combines the likely absence of the parent from the child's life with other factors, such as the degree of criminal involvement of the parent; parents with higher exposure to incarceration will have committed more serious crimes. If the effect of parental incarceration is different for these parents, then the estimates will reflect both the longer sentence and the different incarceration effect. Nonetheless, the estimates are easily interpretable as causal effects for specific subgroups of the population. They can be interpreted as saying that while parental incarceration is positive for most children, the subgroup with parents who have committed more serious crimes may be harmed by it.

One concern with these outcomes is that adult crime measures occur well after the incident of parental incarceration; on average, children in our sample are nine years old when their parent is in court. It may be that incarceration has negative effects in the short-run that either disappear or reverse over a longer time frame. For example, children with incarcerated parents may shift involvement in criminal activity forward in time to commit crimes as a juvenile rather than as an adult. Using juvenile court records in Cuyahoga County, Table 4.19 and Table 4.20 instrument for parental incarceration with judge assignment to test whether having an incarcerated parent is related to appearing in juvenile court or being incarcerated as a juvenile offender. The patterns are even stronger than in the adult court results: having an incarcerated parent decreases

the overall risk of the child appearing in juvenile court or being incarcerated as a juvenile. The effects are large, at between 32-63% of the mean rate of juvenile court appearance and incarceration in the sample. The point estimates are fairly similar across male and female children, as well as between the children of incarcerated mother and fathers, although the standard errors in the subsamples are larger.

As with adult court appearances, the effects are heterogeneous depending on the expected exposure of the child to the parent's incarceration (Table 4.20). For children who are exposed to the shortest stints of parental incarceration, the effects are negative and highly economically significant, decreasing the likelihood of the child being incarcerated or appearing in court by over 76% of the sample mean. For children who are exposed to longer periods of parental incarceration, the sign reverses and we can no longer reject a null effect. Nonetheless, this is consistent with the idea that while short terms of parental incarceration can actually be beneficial for a child, longer stretches do not have such an effect.

Juvenile offenders typically are close to adulthood, and so it may be that the negative effects are concentrated at even earlier ages. If that were the case, we would expect spillovers onto school performance in childhood. Table 4.21 takes a decade of data from Cleveland Metropolitan School District and again instruments for parental incarceration. Although the estimates are noisier, we find no relationship between parental incarceration and child test scores on standardized tests of math and reading or the number of child absences from school.

Children in poor communities may be more vulnerable to shocks such as a losing a parent as they may be lacking in other economic supports. At the same time, parents are on the margin of incarceration may be lower quality caregivers, and the incarceration of a parent could provide an opportunity for the child to shift to a more stable home environment. Table 4.22 breaks

the sample into quartiles based on the level of poverty in the neighborhood in which the child was born.[10] The decrease in criminal justice system involvement for children of incarcerated parents is entirely concentrated among children born in below median wealth neighborhoods, with the strongest effects among those in the bottom quartile. In a future draft of the paper, we will use data from the Department of Human Services to track child removal from their parents, observing whether child was placed with other family members or other foster care arrangement post-parental incarceration. This will help determine the mechanism at work, such as if the positive effects of parental incarceration are indeed related to being placed in a more stable home environment.

One potential concern is that our findings could be driven by differential migration response to parental incarceration. If this migration was outside of Ohio or to counties in Ohio for which we do not measure, it would limit our ability to observe outcomes, potentially biasing estimates. Section A.3.1 addresses this concern using address data from voter records. For individuals who are registered to vote, we can exactly observe whether they have migrated and how far. Children with incarcerated parents are no less likely to be registered to vote in Ohio or have moved outside of these three counties. Those from the poorest birth neighborhoods, where the effects are concentrated, are in fact more likely to be registered in Ohio. Taking the geocoded birth and voter records addresses, we check whether incarceration causes individuals to move further away as adults. If anything, they actually live slightly closer to their birth addresses than those whose parents were not incarcerated, suggesting that out-migration is unlikely to be driving our results.

[10] We take the residential address of the mother from the child's birth certificate and match that to a census block group. Based on American Community Survey data, census block groups are sorted into quartiles by percentage of households with income levels below the poverty line.

3.5.3. Spillovers of Incarceration on Siblings

Although many studies have investigated the spillover effect of parental incarceration, only a handful of correlational studies have looked at the incarceration of siblings (**??**). Compared to parental incarceration, the role of a sibling is less important for caregiving, may free up household resources for other siblings, and may remove a bad influence. On the other hand, of course, sibling incarceration will usually cause an emotional strain or remove a *positive* influence, making the overall effect till ambiguous.

Table 4.23 displays the causal effect of incarceration for siblings. We separate out the effect by relationship: only the same mother, only the same father, same mother, same father, same parents, and all. Broadly, having a sibling who is incarcerated as a result of assignment to a harsher judge reduces an individual's likelihood of criminal justice system involvement and incarceration by around 5 percentage points. These effects are concentrated among children who share the same mother (column (1)), who are presumably more likely to live together and interact regularly. This may also reflect that seeing a sibling's experience in the criminal justice system leads to individuals being "scared straight". More generally, the effects of having an incarcerated sibling are likely to be heterogeneous: in some cases, the social dislocation of having an incarcerated sibling may shift an individual into more criminal activities, while in other cases, it reduces criminal involvement. The second effect clearly dominates the first in this case. Interestingly, siblings who share both parents are not affected by parental incarceration. This is in line with our earlier finding that the benefits of parental incarceration are concentrated among individuals in the bottom quartile of neighborhoods, for whom the rates of single parent families is higher.

Table 4.24 confirms that sibling incarceration leads to the largest reductions in future criminality for those from the poorest stratum. Across the lowest two neighborhood wealth birth quartiles of our sample, incarceration leads to significant, negative reductions in being charged and being incarcerated in the future. For those coming from the two highest neighborhood wealth birth quartiles, we find more mixed evidence on whether sibling incarceration reduces incarceration, with some coefficients turning positive. The measured impacts are concentrated somewhat across siblings sharing mothers, although less so than in the main results. As sibling incarceration is unlikely to affect the overall family structure as the children, it appears that the positive effects of sibling incarceration are likely to be driven by the combined effect of having a disruptive sibling within either family or neighborhood environments that make these influences especially salient.

3.6. Conclusion

In this paper, we generate the first causal estimates of the effects of parental incarceration in the United States and of sibling incarceration in any setting. In contrast to most of the literature we find that parental incarceration leads to decreases in future criminal involvement of children both as juveniles and adults. We find no evidence of effects on academic achievement, and no average effects on teen pregnancy. The positive effects are concentrated among children from poorer neighborhoods, as well as in cases with shorter sentence lengths. As sentence lengths increase, the effects drop to zero. Given the positive extensive margin effects, this suggests that incarceration for short lengths of time is not harmful, but that long spells behind bars are. Indeed, we find that for longer sentences, parental incarceration results in sizeable increases in teen pregnancy.

Given that these results run contrary to most of the existing literature, what mechanisms are behind these results? First, it may be that parental quality is heterogenous and low quality parents are worse than the outside option induced by incarceration (e.g. care by grandparents). This interpretation is bolstered by the fact that our effects are concentrated almost exclusively among defendants living in poor neighborhoods.

Another plausible channel is that children are "scared straight" by the experience of having an incarcerated parent and the increased salience of punishment. Exposure to incarceration may directly cause children to be less likely to commit crime in the future to avoid the criminal justice system. Such a mechanism is consistent with the largest decreases being for short sentences, where the salience of punishment is increased but the reduction in household and parental inputs is relatively small.

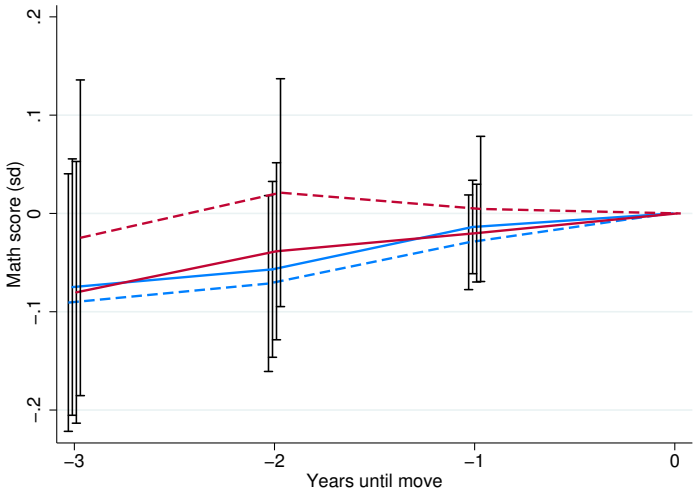
Future drafts of the paper will incorporate new data and new results to parse these mechanisms. For example, we provide suggestive evidence that longer sentences are more harmful to children, but in the future, we will test that directly using variation in sentence length induced by judge assignment. We will also predict parent quality at baseline and see how the effects of incarceration are heterogenous depending on both observable and unobservable measures of parental quality. Future drafts will also include results on additional outcomes such as employment and wages, mental health, substance abuse, and more child educational outcomes, as well as additional results on siblings. These will be helpful in coming to a fuller picture of the costs and benefits of incarceration.

CHAPTER 4

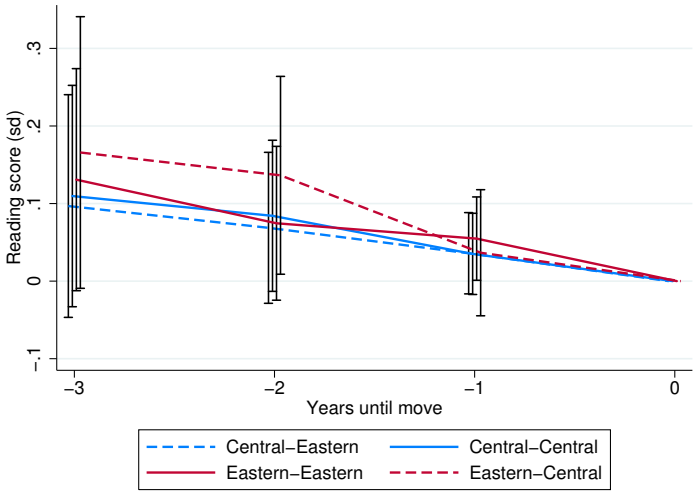
Tables and Figures

Figure 4.1. Pre-move trends in academic outcomes, by mover type

(a) Math scores

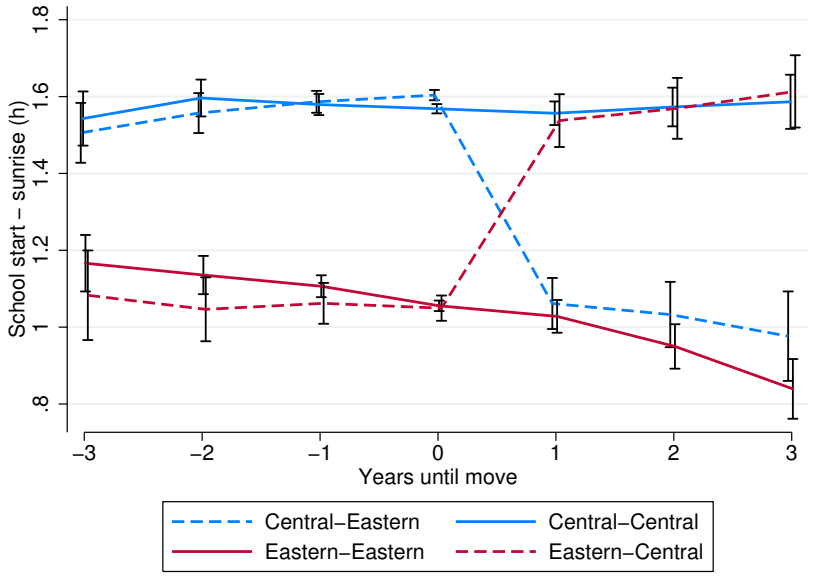


(b) Reading scores



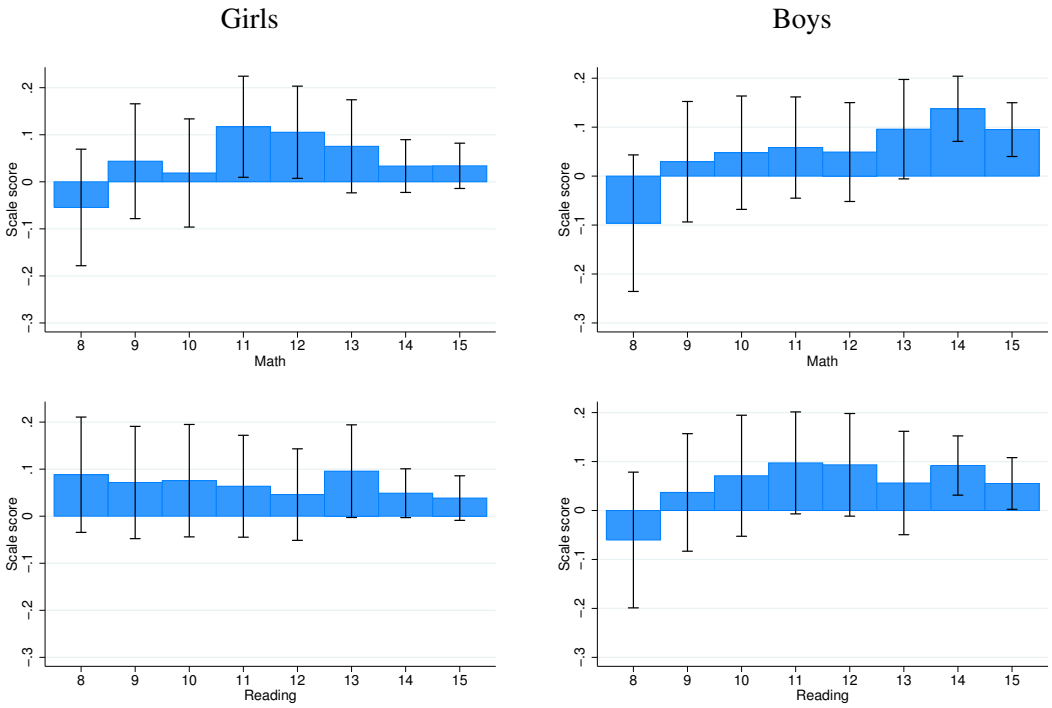
Displays the pre-move achievement trends for the four years leading up to a move of 25 miles or more. Results reported separately for four groups of movers: within CT, within ET, ET to CT, and CT to ET. Coefficients recovered from a regression of test scores on time-until-move dummies, a vector of controls (age-gender dummies, longitude, and school population shares for FRL, male, black, Asian, and Hispanic), and a fixed effect for the period before the move. Standard errors are clustered at the individual level, and included as bars representing 95% confidence intervals.

Figure 4.3. Hours of sunlight before school over move, by mover type



Displays the hours of sunlight before school for four groups: within CT, within ET, ET to CT, and CT to ET. Estimates are from a regression of relative school start time on time relative to move for each mover group, a vector of controls (age-gender dummies, longitude, and school population shares for FRL, male, black, Asian, and Hispanic), and a student-move fixed effect. The year before the move is normalized to be zero; we adjust the level of the coefficients with the group mean of relative start times for one year before the move. Standard errors are clustered at the individual level, and included as bars representing 95% confidence intervals.

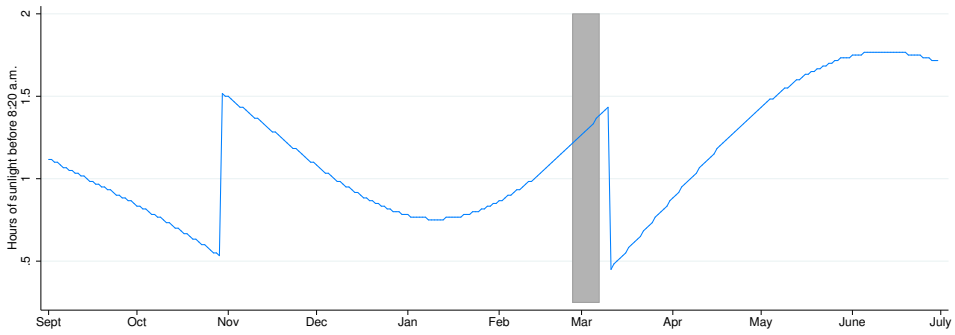
Figure 4.4. Effect of school start times on academic achievement, by age, gender, and subject



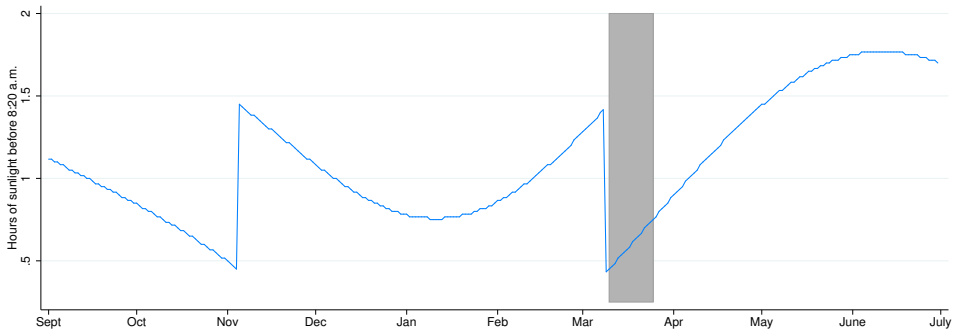
Each subfigure displays the age-gender specific effect of start times on academic achievement. Coefficients are from a regression of scale scores on school start time interacted with age, a vector of controls (age-gender dummies, longitude, and school population shares for FRL, male, black, Asian and Hispanic), and an individual fixed effect. Start time-age interactions are instrumented with time zone-age interactions. Sample is listed in the column headers, dependent variable is noted on the horizontal axis. Standard errors are clustered at the individual level, and included as bars representing 95% confidence intervals.

Figure 4.6. Hours of sunlight before 8:20 a.m. start time, by year with testing periods

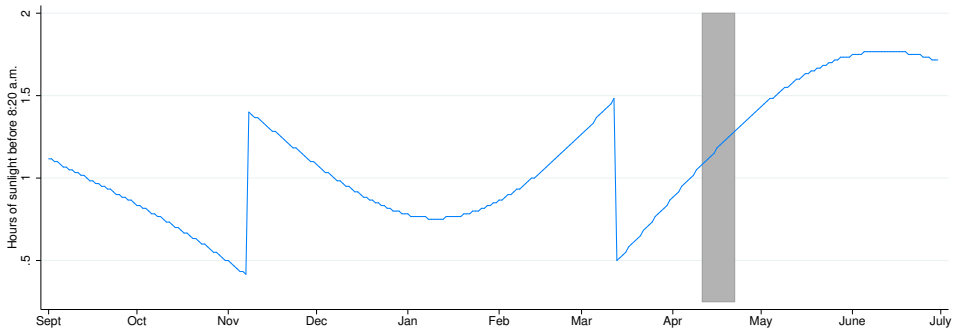
(a) 2000-2007



(b) 2008-2009

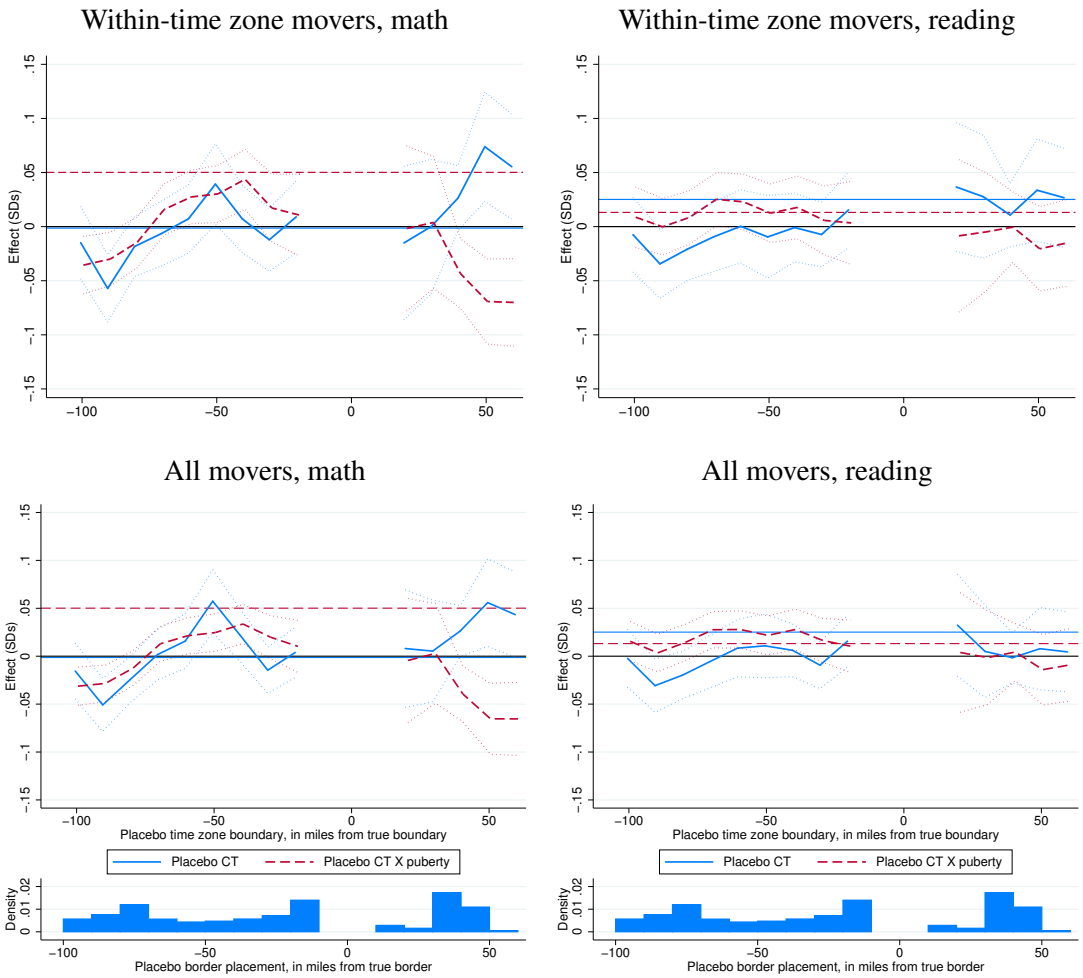


(c) 2011-2013



Amount of sunlight before school and testing dates for a hypothetical school for each of the three testing regimes. School location and opening time chosen to match the average test-day relative start time in ET in 2008. Grey areas represent testing periods. The figures display sunlight for 2007, 2008, and 2011, respectively, but all are archetypes of their era.

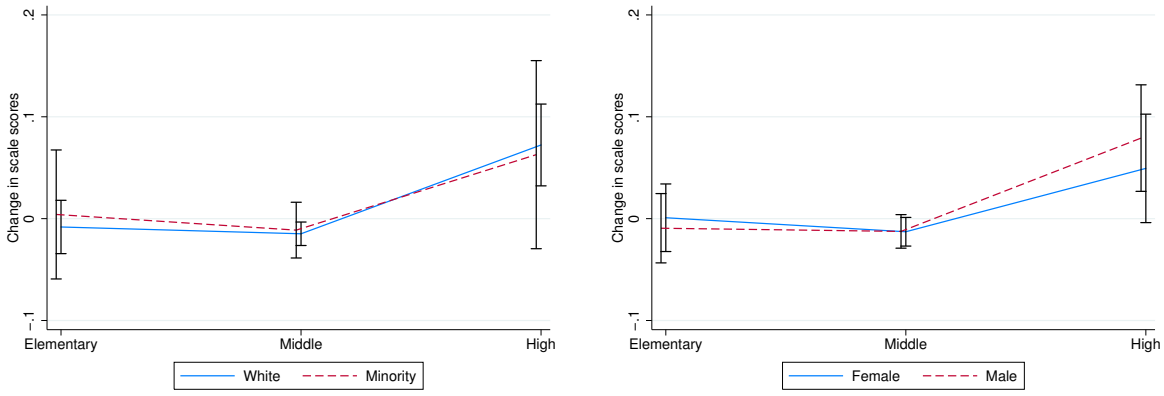
Figure 4.8. Effect of placebo time zones on academic achievement



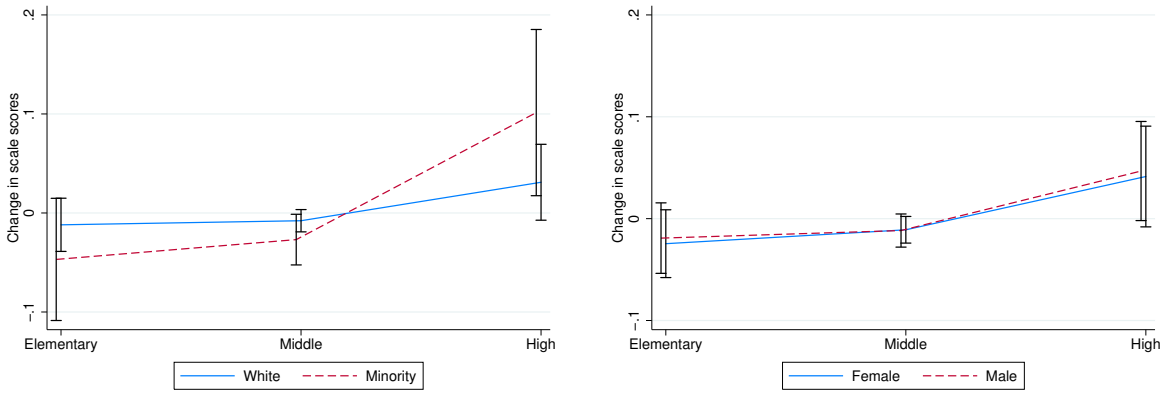
Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Thin horizontal lines represent baseline coefficient estimates. We generate placebo time zones in ten mile increments from the true time zone boundary. Then, placebo coefficients are calculated from individual regressions of the outcome on the true time zone interacted with puberty, and the placebo time zone interacted with puberty. All specifications include age-gender dummies, longitude controls, school demographic means (FRL, male, black, Asian, and Hispanic) and individual fixed effects. Standard errors clustered at the individual level. We display results including and excluding cross-time zone movers. Sample excludes a 25 mile donut around the time zone boundary due to treatment bleed across the boundary.

Figure 4.10. Counterfactual change in test scores, reordered start times

(a) Math Scores



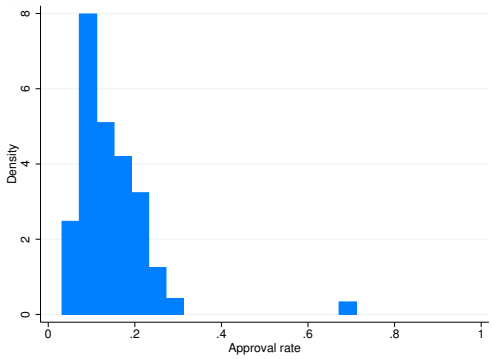
(b) Reading Scores



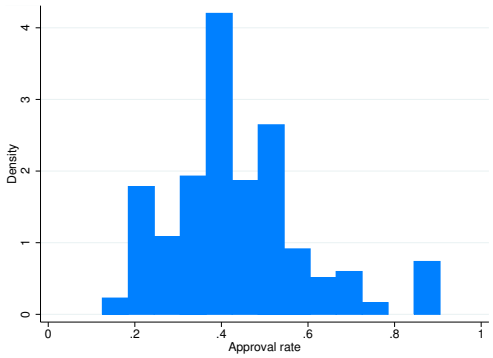
Estimated test score gains under a counterfactual policy where start times are adjusted to be later for older children. Adjustment is conducted by taking the average start time for each school type in each district (elementary, middle, and high), and swapping them between school types so that elementary schools open first, then middle schools, then high schools. We then adjust the level of all school times so that the mean counterfactual district start time is the same as the true mean start time. This results in bell times 22 minutes earlier for elementary schools, 13 minutes earlier for middle schools, and 44 minutes later for high schools. Gains are then calculated by multiplying the changes in start time for each child with the relevant coefficients from Table 4.3. Bars represent 95% confidence intervals.

Figure 4.12. Approval rates by judge

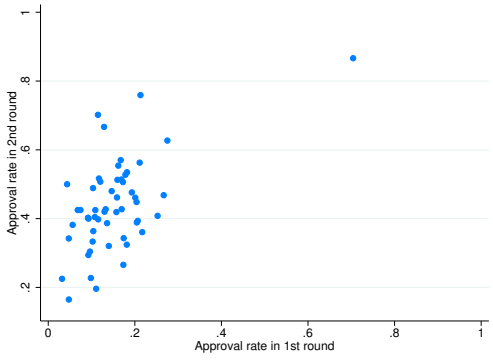
(a) First round



(b) Second round



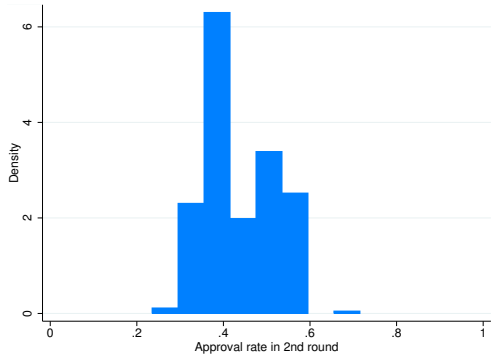
(c) Second vs. first round approval rates



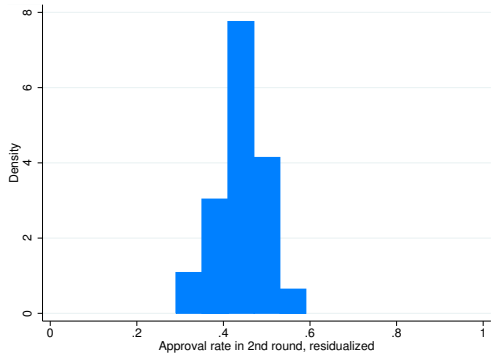
Panel A and B contain histograms of approval rates by judge for the first and second round, respectively. Both are weighted by the number of observations per judge. Panel C contains the scatter plot of judge-level first- and second-round approval rates. The correlation is 0.57, and 0.40 without the outlier.

Figure 4.14. Second-round approval by first-round judge

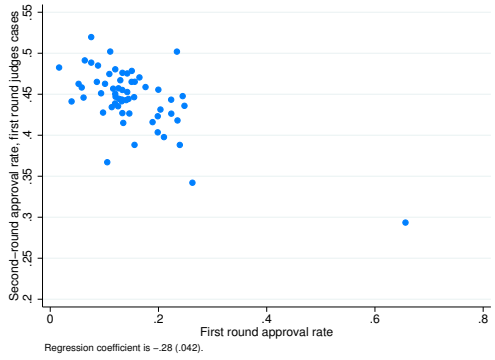
(a) Second-round approval



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



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

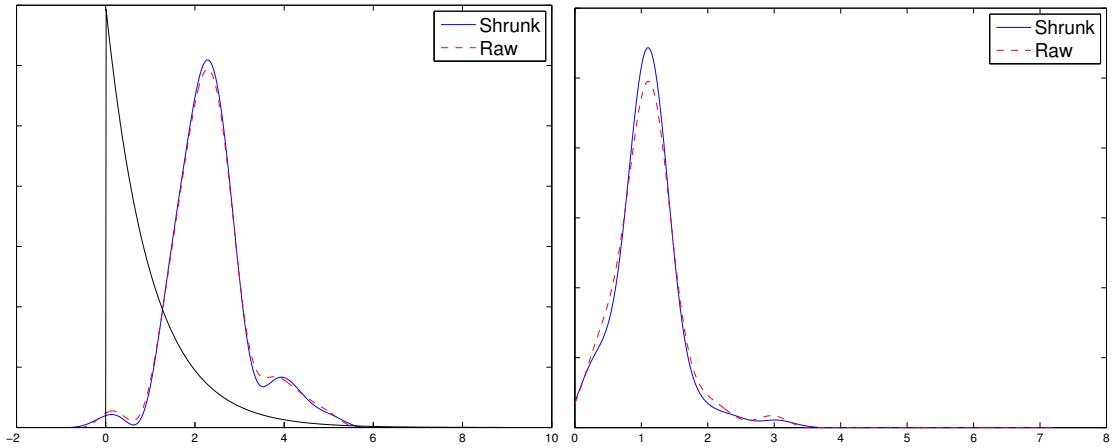


Panel A contains a histogram of second-round approval rates for the cases approved by the first-round judge. Higher approval rates suggest that the first-round judge did a better job of selecting claimants with a high probability of success in the second round. Panel B residualizes out first- and second-round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

Figure 4.16. Distribution of judge coefficients

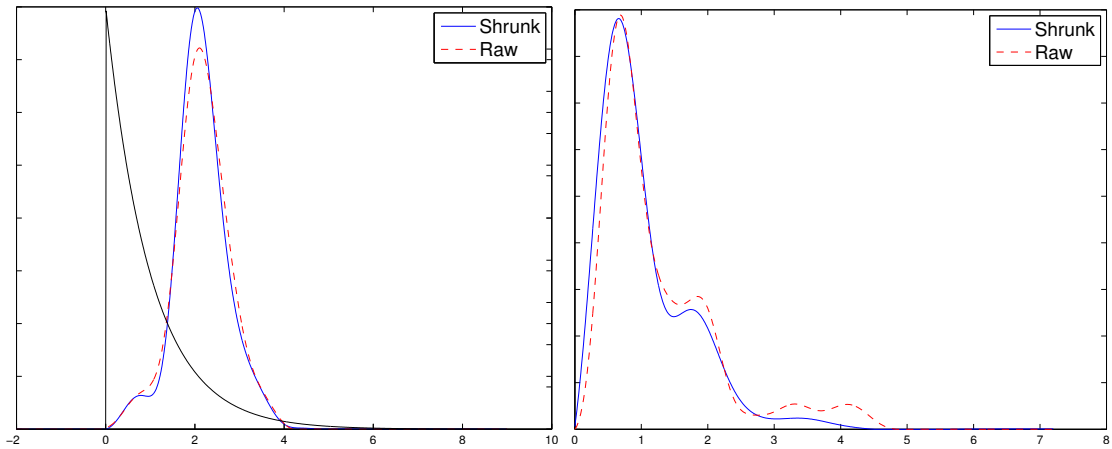
(a) Threshold γ_1, first round

(b) Observational error σ_1, first round



(c) Threshold γ_2, second round

(d) Observational error σ_2, second round



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

Figure 4.18. Model estimates of first- versus second-round approval

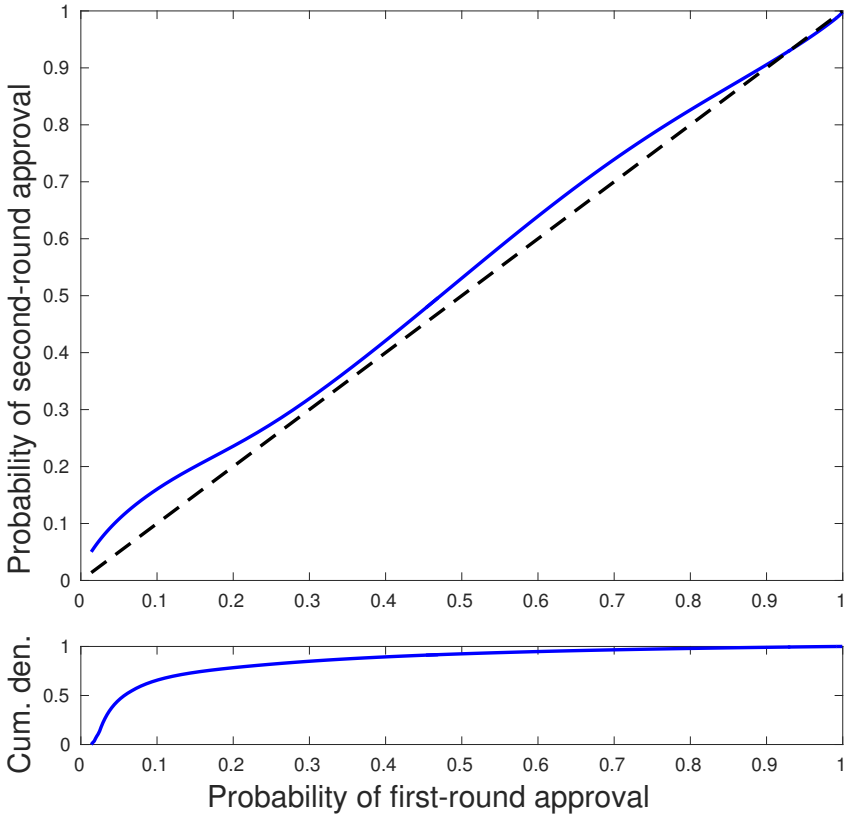
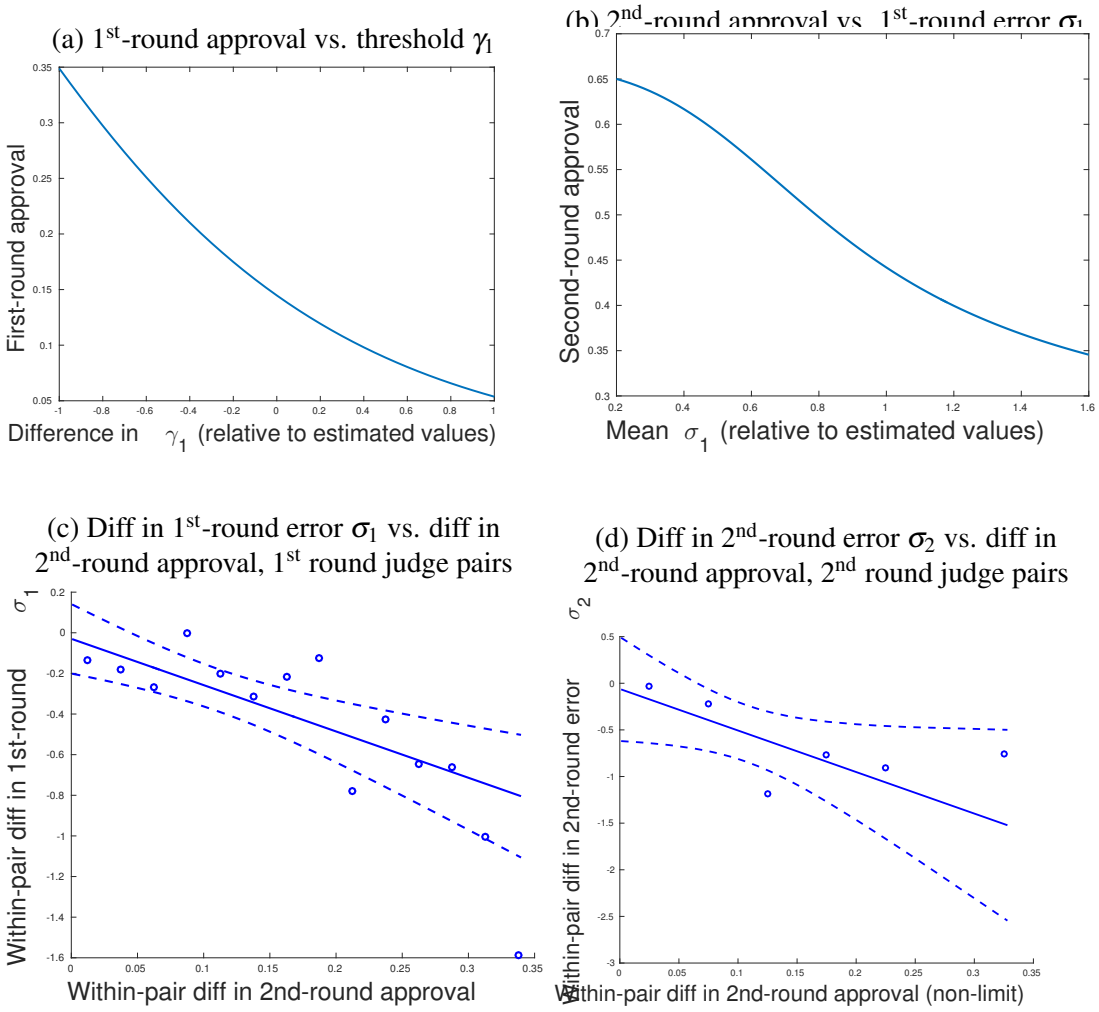


Figure plots first-round approval probability against second-round approval probability conditional on first-round approval for each value of case strength r_i. Secondary graph displays cumulative density of first-round approval. Black dotted comparison line marks out 45°.

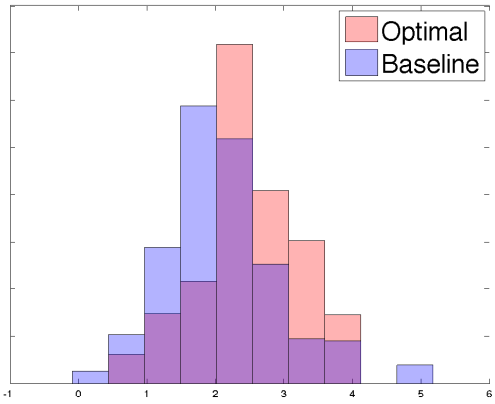
Figure 4.19. Identification intuition



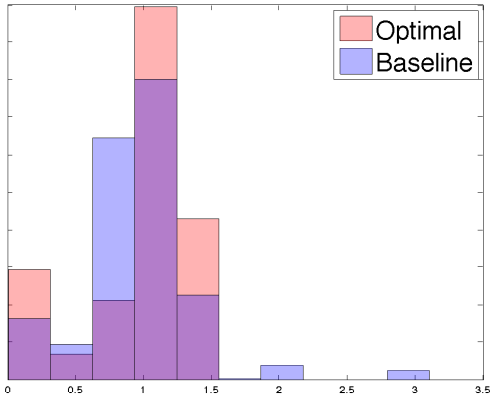
Panel A contains model estimates of the first-round approval probability as a function of deviation of threshold γ_1. Panel B contains model estimates of second-round approval as a function of mean first-round error relative to the estimated value — higher errors make second-round approval less likely. In Panel C, I match pairs of judges with similar first-round approval rates (within 1 percentage point), then display difference in model-estimated judge errors $\widehat{\sigma}_1$. In Panel D, I match pairs of second-round judges with similar approval rates conditional on first-round approval by a high-approving (non-limiting) first-round judge. I then compare the difference in second-round observational error σ_2 as a function of within-pair differences in second-round approval rates conditional on first-round approval by all other judges. See subsubsection 2.2.2.1 for more details.

Figure 4.21. Optimal allocation of judges

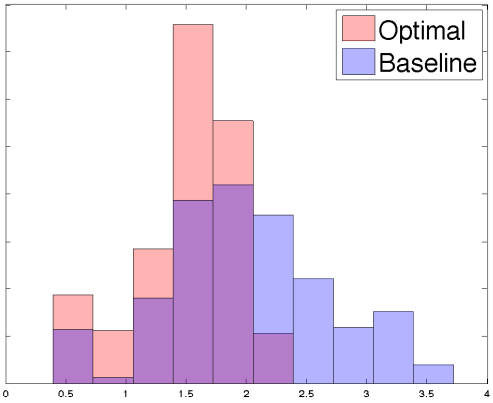
(a) Threshold γ_1, first round



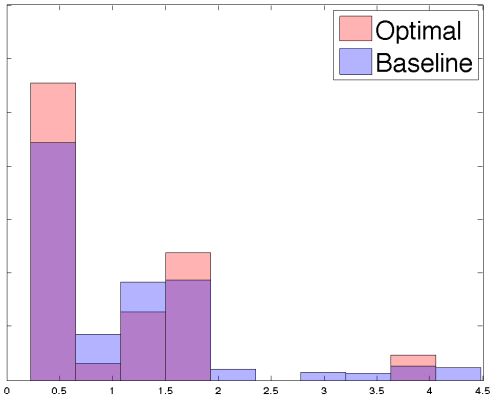
(b) Observational error σ_1, first round



(c) Threshold γ_2, second round

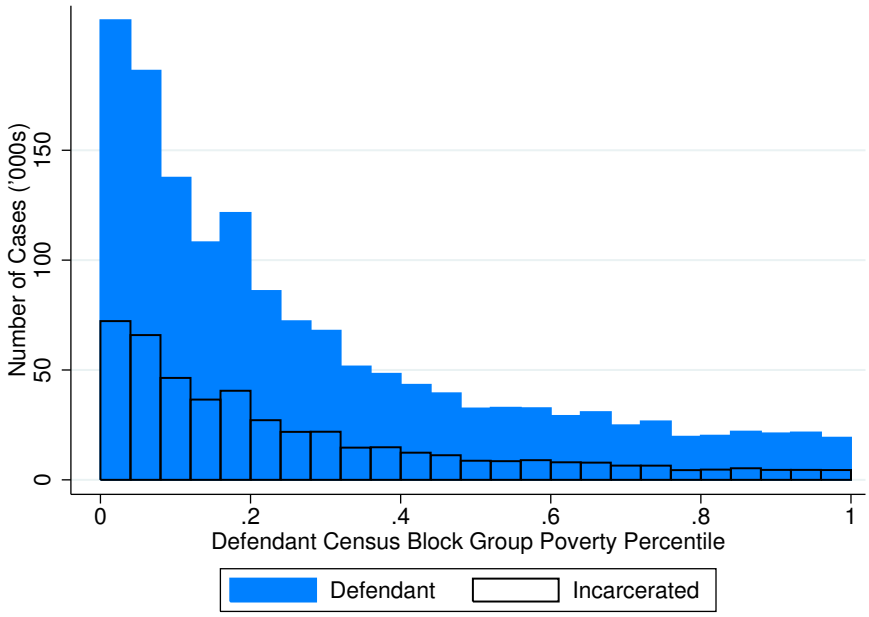


(d) Observational error σ_2, second round



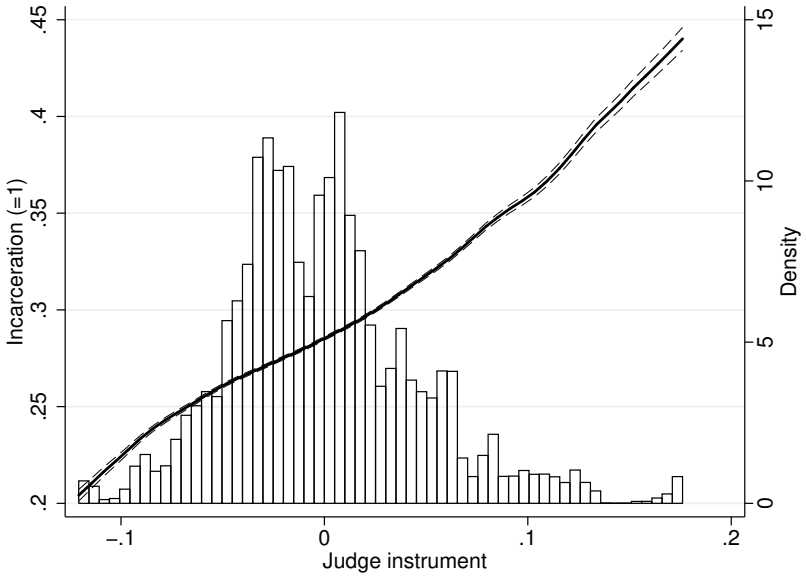
I minimize judge workload requiring that a) no judge works more than she does in the baseline, b) at least as many claimants are approved, and c) the posterior distribution of case strength r_i for approved claimants under the counterfactual first-order stochastically dominates the baseline distribution.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

Figure 4.23. Neighborhood poverty status of defendants



Histogram of poverty percentile of neighborhood at time of arrest for all defendants and incarcerated defendants.

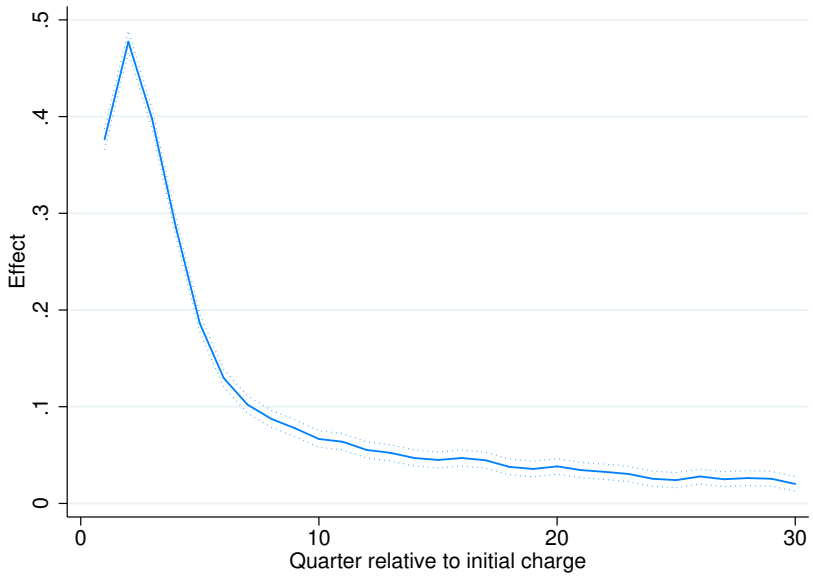
Figure 4.24. First stage of incarceration on judge instrument



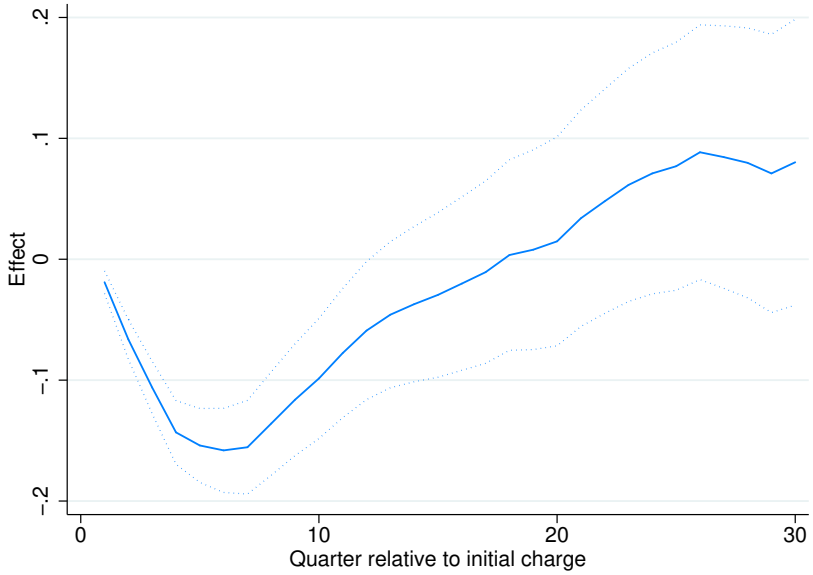
Displays histogram of instrument. Instrument is constructed from leave-out means of judge decisions, residualizing out defendant characteristics and offense. Line is generated by nonparametric regression of incarceration on judge instrument. Robust standard errors are clustered at the judge level.

Figure 4.25. Effect of judge assignment on own incarceration and future crime

(a) Incarcerated (=1)



(b) Number of new charges



Panel A displays coefficients from a quarter-by-quarter regression of present incarceration (on any charge) on the leave-out measure of judge severity. Time is measured since judge assignment. Panel B displays coefficients from a quarter-by-quarter regression of the cumulative number of new charges since judge assignment on leave-out judge severity. Standard errors are clustered at the judge level, and included as dotted lines representing 95% confidence intervals.

Table 4.1. Sample characteristics, Florida panhandle movers

	Panhandle (1)	Movers (2)	CT-ET (3)	ET-CT (4)	Difference (3)-(4)
Panel A: School characteristics					
FRL (fraction)	0.54	0.55	0.56	0.56	0.000
	[0.27]	[0.24]	[0.21]	[0.30]	(0.038)
Male (fraction)	0.51	0.51	0.51	0.51	0.003
	[0.02]	[0.02]	[0.03]	[0.03]	(0.004)
Black (fraction)	0.25	0.26	0.20	0.37	-0.168***
	[0.27]	[0.28]	[0.22]	[0.47]	(0.057)
Hispanic (fraction)	0.04	0.03	0.03	0.03	-0.008
	[0.04]	[0.04]	[0.02]	[0.07]	(0.008)
Asian (fraction)	0.02	0.02	0.01	0.01	0.004
	[0.02]	[0.02]	[0.02]	[0.02]	(0.003)
District Grade 3 math scores (SD)	0.11	0.11	0.12	0.08	0.039
	[0.22]	[0.25]	[0.21]	[0.34]	(0.043)
District Grade 3 reading scores (SD)	0.15	0.15	0.17	0.09	0.084**
	[0.22]	[0.23]	[0.17]	[0.34]	(0.041)
District Grade 3 absentee rates	4.54	4.48	4.39	4.74	-0.358
	[0.77]	[1.13]	[1.70]	[1.40]	(0.227)
1999 median income by zip, logged	10.67	10.64	10.59	10.62	-0.036
	[0.27]	[0.26]	[0.26]	[0.36]	(0.051)
Student/teacher ratio	15.43	15.72	15.40	15.80	-0.400
	[1.20]	[1.40]	[2.15]	[1.71]	(0.273)
Charter school (fraction)	0.02	0.01	0.01	0.02	-0.015
	[0.12]	[0.07]	[0.05]	[0.14]	(0.017)
Urban (fraction)	0.27	0.24	0.18	0.27	-0.086
	[0.48]	[0.47]	[0.49]	[0.63]	(0.084)
Panel B: Individual characteristics					
FRL (=1)	0.55	0.66	0.67	0.69	-0.017
	[0.50]	[0.47]	[0.47]	[0.46]	(0.025)
Male (=1)	0.52	0.51	0.51	0.52	-0.003
	[0.50]	[0.50]	[0.50]	[0.50]	(0.026)
Black (=1)	0.26	0.26	0.25	0.26	-0.008
	[0.44]	[0.44]	[0.44]	[0.44]	(0.023)
Hispanic (=1)	0.04	0.04	0.02	0.03	-0.008
	[0.19]	[0.20]	[0.15]	[0.18]	(0.009)
Asian (=1)	0.02	0.01	0.01	0.01	0.002
	[0.13]	[0.11]	[0.10]	[0.09]	(0.005)
Math score (SD)	0.11	0.02	-0.06	0.00	-0.064
	[0.96]	[0.92]	[0.88]	[0.88]	(0.047)
Reading score (SD)	0.15	0.07	0.00	0.00	0.003
	[0.97]	[0.93]	[0.90]	[0.93]	(0.048)
Absentee rate	4.52	5.60	5.44	6.46	-1.026***
	[4.44]	[5.18]	[5.16]	[5.60]	(0.325)
Observations	186,278	13,788	713	726	

Sample is all third graders in the panhandle. Categorical variables are reported as 0-1. Absentee rate is reported as the percentage (0-100) of days missed in the school year to ease interpretation. Standard deviations in square brackets. Standard errors in parentheses and clustered at the school level in Panel A, unclustered in Panel B. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.2. Academic and behavioral outcomes on start time, with student fixed effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: First stage, relative start time (hours)							
CT (=1)	0.471***	0.345***	0.424***	0.415***	0.346***	0.424***	0.415***
	(0.016)	(0.021)	(0.020)	(0.020)	(0.021)	(0.020)	(0.020)
CT X Puberty	0.264***	0.265***	0.306***	0.265***	0.265***	0.306***	0.265***
	(0.012)	(0.012)	(0.011)	(0.011)	(0.012)	(0.011)	(0.011)
P(CT+CT X puberty=0)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Panel B: IV estimates, math test scores (SDs) on relative start time							
Start time - sunrise (h)	-0.063**	0.014	0.020	0.010	0.012	0.020	0.009
	(0.026)	(0.041)	(0.036)	(0.035)	(0.041)	(0.036)	(0.035)
Start time X puberty	0.099***	0.074***	0.058***	0.074***	0.073***	0.057***	0.073***
	(0.018)	(0.020)	(0.021)	(0.019)	(0.020)	(0.021)	(0.019)
P(Start+Start X puberty=0)	0.042	0.002	0.001	0.001	0.002	0.001	0.001
Cragg-Donald F-stat	1101.18	404.14	588.90	541.51	405.14	588.76	542.01
Panel C: IV estimates, reading test scores (SDs) on relative start times							
Start time - sunrise (h)	0.064**	0.088**	0.081**	0.061*	0.087**	0.081**	0.061*
	(0.027)	(0.041)	(0.037)	(0.036)	(0.041)	(0.037)	(0.036)
Start time X puberty	-0.005	-0.014	-0.023	-0.005	-0.013	-0.023	-0.004
	(0.018)	(0.021)	(0.022)	(0.020)	(0.021)	(0.022)	(0.020)
P(Start+Start X puberty=0)	0.000	0.004	0.008	0.014	0.004	0.008	0.014
Cragg-Donald F-stat	1230.00	485.69	637.13	618.88	486.65	637.22	619.26
Panel D: IV estimates, absence rate (%) on relative start times							
Start time - sunrise (h)	-0.937***	-1.885***	-0.696	-0.856*	-1.860***	-0.718	-0.869*
	(0.361)	(0.594)	(0.476)	(0.487)	(0.590)	(0.474)	(0.485)
Start time X puberty	0.481**	0.846***	0.365	0.443*	0.857***	0.395	0.469*
	(0.245)	(0.295)	(0.286)	(0.268)	(0.294)	(0.285)	(0.268)
P(Start+Start X puberty=0)	0.062	0.008	0.264	0.206	0.010	0.274	0.219
Cragg-Donald F-stat	689.75	273.69	425.19	383.57	274.18	425.38	383.62
Longitude	No	Yes	Yes	Yes	Yes	Yes	Yes
District quality	No	No	Yes	No	No	Yes	No
School quality	No	No	No	Yes	No	No	Yes
Time since move	No	No	No	No	Yes	Yes	Yes

Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Absentee rate is the fraction of days the child missed school. Start time and its interaction with puberty are instrumented by time zone. Sample is all children who moved. All specifications include age-gender dummies and individual fixed effects. Sample size is fixed within panels: 34,018 students and 115,778 student-years in Panel A, 24,768 students and 99,835 student-years in Panel B, 25,191 students and 104,791 student-years in Panel C, and 15,906 students and 66,263 student-years in Panel D. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.3. Academic and behavioral outcomes on start time, by group with student fixed effects

	White	Non-white	Non-FRL	FRL	Male	Female
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Math test scores (SDs)						
Start time - sunrise (h)	0.021	-0.017	0.045	-0.015	0.027	-0.008
	(0.039)	(0.095)	(0.055)	(0.046)	(0.050)	(0.050)
Start time X puberty	0.072***	0.098**	0.102***	0.063**	0.076***	0.072***
	(0.022)	(0.046)	(0.032)	(0.025)	(0.027)	(0.028)
P(Start+Start X puberty=0)	0.000	0.182	0.000	0.137	0.003	0.069
Cragg-Donald F-stat	459.66	84.63	177.22	373.97	263.79	277.79
Number of students	17013	7755	10052	14716	12380	12388
Observations	70535	29300	40140	59695	49436	50399
Panel B: Reading test scores (SDs)						
Start time - sunrise (h)	0.034	0.135	0.072	0.056	0.055	0.072
	(0.040)	(0.092)	(0.056)	(0.047)	(0.051)	(0.050)
Start time X puberty	0.006	-0.003	-0.028	0.006	0.006	-0.018
	(0.024)	(0.046)	(0.035)	(0.025)	(0.028)	(0.029)
P(Start+Start X puberty=0)	0.113	0.018	0.215	0.037	0.060	0.101
Cragg-Donald F-stat	516.36	100.07	221.60	407.29	289.00	333.87
Number of students	17264	7927	10284	14907	12560	12631
Observations	73872	30919	42458	62333	51752	53039
Panel C: Absence rate (%)						
Start time - sunrise (h)	-0.357	-2.012	-1.094	-0.619	-0.564	-1.277*
	(0.531)	(1.312)	(0.737)	(0.625)	(0.622)	(0.752)
Start time X puberty	-0.193	1.723***	0.298	0.533	0.201	0.794**
	(0.324)	(0.622)	(0.411)	(0.343)	(0.379)	(0.377)
P(Start+Start X puberty=0)	0.123	0.720	0.089	0.840	0.379	0.346
Cragg-Donald F-stat	320.62	58.76	116.36	270.00	193.18	190.14
Number of students	10613	5293	6383	9523	8019	7887
Observations	45654	20609	26483	39780	32994	33269

Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Absentee rate is the percent of days the child missed school. Start time and its interaction with puberty are instrumented by time zone. Sample is all children who moved more than 25 miles. All specifications include age-gender dummies, longitude controls, school demographic means (FRL, male, black, Asian, and Hispanic) and individual fixed effects. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.4. Persistence in effects of relative start time on student outcomes, with student fixed effects

	Math score (in SD)		Reading score (in SD)	
	(1)	(2)	(3)	(4)
Start time - sunrise (h) (prepubescent)	0.009	0.007	0.061*	0.052
	(0.035)	(0.036)	(0.036)	(0.036)
Start X moved two years ago (pre)		0.002		0.011
		(0.009)		(0.009)
Start X moved 3+ years ago (pre)		-0.011		-0.005
		(0.012)		(0.012)
Start time - sunrise (h) (pubescent)	0.082***	0.087***	0.057**	0.048**
	(0.025)	(0.026)	(0.023)	(0.024)
Start X moved two years ago (pub)		-0.016***		-0.004
		(0.006)		(0.006)
Start X moved 3+ years ago (pub)		-0.020***		0.010
		(0.007)		(0.007)
P[Start (pre) = Start (pub)]	0.000	0.000	0.826	0.861
P[Start (pre) = Start (pub), long run]		0.000		0.577
Cragg-Donald F-stat	542.01	107.47	619.26	124.19
Number of students	24,768	24,768	25,191	25,191
Observations	99,835	99,835	104,791	104,791

Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Start time and its interaction with puberty are instrumented by time zone and the interaction of time zone and puberty. Sample is all children who moved more than 25 miles. All specifications include age-gender dummies, longitude controls, school demographic means (FRL, male, black, Asian, and Hispanic) and individual fixed effects. Standard errors in parentheses and clustered at the individual level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.5. Academic outcomes, for testing before and after DST

	Math (SD)		Reading (SD)	
	(1)	(2)	(3)	(4)
Start time - sunrise (h) (prepubescent)	0.030		0.056	
	(0.038)		(0.038)	
Start time - sunrise (h) (pubescent)	0.096***		0.060**	
	(0.027)		(0.024)	
Start time X prepubescent, late test time		0.022		0.049
		(0.039)		(0.039)
Start time X prepubescent, early test time		0.071		0.096**
		(0.046)		(0.047)
Start time X adolescent, late test time		0.095***		0.045*
		(0.030)		(0.026)
Start time X adolescent, early test time		0.096***		0.104***
		(0.025)		(0.026)
Era X puberty controls	No	Yes	No	Yes
P[Early = late test (Prepub)]		0.165		0.192
P[Early = late test (Adol)]		0.967		0.001
Cragg-Donald F-stat	468.563	229.684	542.050	269.539
Number of students	23,618	23,618	24,152	24,152
Observations	89,707	89,707	94,515	94,515

Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Start time and its interactions are instrumented by time zone and the interaction of time zone and interactions. Sample is all children who moved more than 25 miles. All specifications include age-gender dummies, longitude controls, school demographic means (FRL, male, black, Asian, and Hispanic), time since move dummies, and individual fixed effects. Sample includes years 2000-2013 excluding 2010, when testing took place over the DST time change. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.6. Randomization

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

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

Table 4.7. Second-round approval on mean approval rate of first-round judge

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

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

Table 4.8. Placebo tests and relevance for regressors, with judge fixed effects

	Predicted approval		Actual approval	
	(1)	(2)	(3)	(4)
Panel A: First round				
End of week	0.000	0.000	-0.008***	-0.007***
	(0.000)	(0.000)	(0.002)	(0.002)
Observations	58604	58604	58604	58604
Panel B: Second round				
End of week	0.001	0.001	-0.022*	-0.022*
	(0.001)	(0.001)	(0.012)	(0.012)
Noon hearing	-0.001	-0.001	-0.078***	-0.075***
	(0.001)	(0.001)	(0.022)	(0.023)
Controls	No	Yes	No	Yes
Observations	8,446	8,446	8,446	8,446

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

Table 4.9. Second-round outcome on model approval probability and judge-pair FEs

	Judge-pair round FEs		Judge-pair FEs	
	(1)	(2)	(3)	(4)
Model approval probability	0.945***	0.938***	0.967***	0.962***
	(0.146)	(0.169)	(0.0463)	(0.0470)
Model controls	No	Yes	No	Yes
Mean approval	0.44	0.44	0.44	0.44
F-stat for judge pairs	1.01	1.02	0.99	0.99
P-value	0.350	0.344	0.619	0.615
Bootstrap p-value	0.694	0.693	0.810	0.809
SD of judge-pair EB means	0.004	0.004	0.004	0.004
Observations	8,196	8,196	8,196	8,196

Regresses second-round approval on model-predicted likelihood of approval and judge-pair fixed effects. Left two columns construct judge-pair FEs accounting for order of assignment; right two columns ignore this distinction. Model controls include office of origination, pre-post 2002, and an end-of-week and noon hearing dummy for the second-round hearing. Standard errors clustered at the judge level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.10. First-round judge consistency by experience and workload

	(1)	(2)	(3)	(4)	(5)
Coefficients ψ affecting judge inconsistency σ_1					
Experience $>$ 1 year	-0.774^{***}	-0.688^{***}		-0.337^{***}	-0.460^{***}
	(0.096)	(0.255)		(0.119)	(0.144)
Experience $>$ 5 years	-0.290^{***}	-0.237^{***}		0.047	-0.053
	(0.046)	(0.079)		(0.080)	(0.383)
Experience $>$ 10 years	-0.536^{***}	-0.443^{**}		-0.476^{***}	-0.509^{***}
	(0.175)	(0.205)		(0.100)	(0.113)
Log monthly caseload			0.197^{***}	0.239^{***}	
			(0.009)	(0.023)	
Log caseload ($\leq$ 5 yrs exp)					0.179^{***}
					(0.040)
Log caseload ($>$ 5 yrs exp)					0.062
					(0.100)
Second-round experience control	Yes	Yes	No	Yes	Yes
Career number of cases	No	Yes	No	No	No

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

Table 4.11. Model coefficients on survey responses

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Threshold γ_1 (mean=2.26, SD=.87)						
Favorability, SD	-0.144	-0.237***			-0.055	-0.186*
	(0.104)	(0.086)			(0.114)	(0.102)
Inconsistency, SD			0.316***	0.212***	0.304***	0.134*
			(0.046)	(0.055)	(0.052)	(0.071)
Respondent FE	No	Yes	No	Yes	No	Yes
Observations	182	182	182	182	182	182
Panel B: Threshold γ_2 (mean=2.19, SD=.99)						
Favorability, SD	-0.280**	-0.402***			-0.264***	-0.420***
	(0.107)	(0.080)			(0.097)	(0.090)
Inconsistency, SD			0.121	0.104	0.058	-0.053
			(0.078)	(0.093)	(0.069)	(0.079)
Respondent FE	No	Yes	No	Yes	No	Yes
Observations	182	182	182	182	182	182
Panel C: Inconsistency σ_1 (mean=1.89, SD=2.17)						
Favorability, SD	0.082	0.020			0.152***	0.092
	(0.060)	(0.099)			(0.044)	(0.094)
Inconsistency, SD			0.184***	0.163***	0.224***	0.194***
			(0.047)	(0.056)	(0.049)	(0.061)
Respondent FE	No	Yes	No	Yes	No	Yes
Observations	182	182	182	182	182	182

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

Table 4.12. Inconsistency before and after judge selection reform

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

Estimated with Hanushek (1974) correction for estimated dependent variable. Dependent variable is consistency σ_{j1}, which is estimated from decision model $\mathbb{1}[r_i > X_{ijs}\beta_s + \gamma_{js} + \widetilde{\varepsilon}_{ijs}(W_{ijs})]$, $\widetilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\widetilde{\sigma}_{js} + W_{ijs}\psi_s)$. In the right-hand panel, β_s and ψ_s include dummies for more than 1, 5, and 10 years of experience. Robust standard errors in parentheses and clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.13. Defendant, judge, and court characteristics, by county (Common Pleas)

	Cuyahoga	Franklin	Hamilton
Black	0.69	0.50	0.63
	[0.48]	[0.51]	[0.50]
Male	0.82	0.79	0.82
	[0.37]	[0.40]	[0.38]
Age	32.73	32.36	29.86
	[10.89]	[10.66]	[10.91]
Neighborhood Median Income	33,276.64	37,972.48	34,827.29
	[19,801.40]	[21,054.95]	[21,970.50]
Perc of Neighborhood Below Poverty Line	0.33	0.32	0.34
	[0.20]	[0.20]	[0.22]
Violent crime	0.12	0.10	0.08
	[0.33]	[0.30]	[0.28]
Property crime	0.31	0.43	0.34
	[0.48]	[0.53]	[0.50]
Drug crime	0.38	0.26	0.32
	[0.51]	[0.48]	[0.49]
First Time Offender	0.53	0.39	0.46
	[0.49]	[0.49]	[0.50]
Sentenced to incarceration (=1)	0.31	0.43	0.48
	[0.46]	[0.50]	[0.50]
Sentence length (in days)	246.11	316.83	348.38
	[895.42]	[1,077.74]	[1,144.52]
Sentence if incarcerated (in days)	790.28	783.19	896.49
	[1,488.80]	[1,566.58]	[1,687.65]
Observations	233,299	110,820	103,418

Sample is all defendants in Cuyahoga, Hamilton and Franklin counties common pleas courts. Standard deviations in square brackets and standard errors in parentheses.

Table 4.14. Defendant, judge, and court characteristics by parental status (Municipal and Common Pleas)

	Parent	Non-parent	Difference
Black	0.54	0.55	-0.02***
	[0.49]	[0.47]	(0.001)
Male	0.68	0.81	-0.13***
	[0.45]	[0.37]	(0.001)
Age	31.12	32.48	-1.66***
	[9.14]	[10.82]	(0.022)
Neighborhood Median Income	36,993.16	36,670.14	199.65***
	[21546.38]	[21,388.21]	(53.646)
Perc of Neighborhood Below Poverty Line	0.32	0.32	-0.01***
	[0.20]	[0.20]	(0.000)
Assigned public defender (=1)	0.32	0.31	-0.00
	[0.46]	[0.44]	(0.001)
Violent crime	0.16	0.15	0.01***
	[0.36]	[0.34]	(0.001)
Property crime	0.25	0.26	-0.01***
	[0.43]	[0.42]	(0.001)
Drug crime	0.22	0.24	-0.02***
	[0.41]	[0.42]	(0.001)
Number of previous charges	2.19	2.32	-0.20***
	[4.05]	[4.76]	(0.010)
Number of previous Common Pleas charges	0.37	0.42	-0.07***
	[1.06]	[1.23]	(0.003)
Number of previous Municipal charges	1.68	1.74	-0.14***
	[3.70]	[4.60]	(0.010)
Sentenced to incarceration (=1)	0.26	0.30	-0.04***
	[0.43]	[0.43]	(0.001)
Sentence length (in days)	116.68	158.77	-28.20***
	[593.09]	[717.43]	(1.347)
Sentence if incarcerated (in days)	216.44	246.08	-11.89***
	[456.51]	[520.05]	(0.859)
Observations	305,416	669,679	

Sample is all defendants in Cuyahoga, Hamilton, and Franklin counties courts. Includes controls for county. Note that this includes both common pleas and municipal court defendants, so values differ from those in table 1, which is solely common pleas. Standard deviations in square brackets and standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.15. First stage for group versus overall, leave-out mean

	(1) All	(2) Parent	(3) Mother	(4) Father	(5) Black	(6) Drugs	(7) Age ≤ 30	(8) Age ≥ 30	(9) Severity tercile 1	(10) Severity tercile 2	(11) Severity tercile 3
Leave-out mean	0.888***	0.877***	0.810***	0.902***	0.915***	1.021***	0.888***	0.888***	0.712***	1.010***	0.870***
	(0.0243)	(0.0282)	(0.0470)	(0.0310)	(0.0299)	(0.0665)	(0.0306)	(0.0272)	(0.0390)	(0.0367)	(0.0303)
Observations	975,446	305,188	97,629	207,559	534,107	129,339	484,484	490,962	327,593	316,783	315,631
Ratio relative to overall		0.988	0.913	1.016	1.030	1.150	1.000	1.000	0.802	1.138	0.980
		(.042)	(.059)	(.045)	(.044)	(.081)	(.044)	(.041)	(.049)	(.052)	(.043)

Dependent variable in header. Controls include year-court fixed effects. Standard errors in parentheses and clustered at the crime level. Ratio standard erorrs calculated via the delta method. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.16. Placebo tests for judge severity

	(1) Drugs	(2) Charge sen.	(3) Log sen.	(4) Male	(5) Black	(6) Age	(7) SNAP Perc	(8) Median Income
Incarceration leave-out mean	-0.00652	-0.891	-0.0168	0.00807	0.0121	-0.179	0.00530	-442.6
	(0.0109)	(3.425)	(0.0344)	(0.00917)	(0.0105)	(0.241)	(0.00352)	(420.3)
Dependent mean	0.24	76.53	2.59	0.77	0.55	32.05	0.31	36773.38
Observations	961782	960020	948239	966087	973623	975447	805291	775737

Dependent variable in header. Controls include year-court fixed effects. Standard errors in parentheses and clustered at the judge level at the crime level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.17. Effect of incarceration on child outcomes

	All	Boys	Girls	All	Boys	Girls
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: In court						
Parent incarcerated (=1)	-0.028	-0.047	-0.001			
	(0.021)	(0.045)	(0.028)			
Mother incarcerated (=1)				-0.108***	-0.124**	-0.086*
				(0.034)	(0.059)	(0.049)
Father incarcerated (=1)				0.029	0.006	0.061**
				(0.021)	(0.043)	(0.028)
Dependent mean	0.241	0.297	0.175	0.241	0.297	0.175
Observations	158315	85650	72665	158315	85650	72665
Panel B: Incarcerated						
Parent incarcerated (=1)	-0.031*	-0.058*	0.008			
	(0.017)	(0.033)	(0.012)			
Mother incarcerated (=1)				-0.072***	-0.109**	-0.023
				(0.028)	(0.050)	(0.021)
Father incarcerated (=1)				-0.002	-0.024	0.030**
				(0.014)	(0.027)	(0.015)
Dependent mean	0.083	0.126	0.032	0.083	0.126	0.032
Observations	158276	85637	72639	158276	85637	72639
Panel C: Pregnant before 18						
Parent incarcerated (=1)	0.008	-0.001	0.019			
	(0.012)	(0.008)	(0.022)			
Mother incarcerated (=1)				0.004	-0.001	0.009
				(0.020)	(0.014)	(0.036)
Father incarcerated (=1)				0.011	-0.001	0.027
				(0.011)	(0.008)	(0.021)
Dependent mean	0.043	0.013	0.078	0.043	0.013	0.078
Observations	133908	70575	63333	133908	70575	63333

Incarceration instrumented by judge leave-out incarceration rate. All specifications include year X court fixed effects, and indicators for previous court appearances and incarcerations. Standard errors clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.18. Effect of incarceration on child outcomes

	In court (=1)			Incarcerated (=1)			Teen pregnancy (=1)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Panel A: All parents								
	All	Boys	Girls	All	Boys	Girls	All	Boys	Girls
Parent incarcerated (=1)	-0.067***	-0.093	-0.033	-0.049**	-0.108***	0.030*	-0.011	-0.003	-0.023
	(0.024)	(0.060)	(0.038)	(0.019)	(0.040)	(0.015)	(0.017)	(0.010)	(0.033)
Incarcerated X exposure	0.066**	0.080*	0.051	0.031*	0.086**	-0.040*	0.038**	0.006	0.076*
	(0.029)	(0.048)	(0.033)	(0.017)	(0.038)	(0.021)	(0.018)	(0.010)	(0.040)
10th percentile	-.063***	-.089	-.03	-.048**	-.104***	.028*	-.009	-.002	-.019
	(.024)	(.058)	(.037)	(.019)	(.039)	(.015)	(.016)	(.01)	(.032)
50th percentile	-.053**	-.076	-.022	-.042**	-.089**	.021*	-.003	-.001	-.007
	(.021)	(.053)	(.033)	(.018)	(.036)	(.013)	(.014)	(.009)	(.027)
90th percentile	.006	-.006	.023	-.015	-.014	-.014	.03**	.004	.061**
	(.026)	(.043)	(.028)	(.018)	(.035)	(.016)	(.013)	(.009)	(.028)
Dependent mean	0.241	0.297	0.175	0.241	0.126	0.032	0.043	0.013	0.078
Observations	155497	84117	71380	155460	84104	71356	131528	69312	62216
	Panel B: By gender of parent								
Mother incarcerated (=1)	-0.177***	-0.207***	-0.146**	-0.098***	-0.156***	-0.022	-0.024	-0.007	-0.048
	(0.041)	(0.070)	(0.071)	(0.032)	(0.060)	(0.028)	(0.029)	(0.017)	(0.057)
Mother incarcerated X exposure	0.095*	0.069	0.152*	0.055	0.099**	-0.022	0.043*	-0.005	0.111*
	(0.052)	(0.057)	(0.080)	(0.034)	(0.048)	(0.027)	(0.025)	(0.017)	(0.063)
Father incarcerated (=1)	0.015	-0.015	0.045	-0.011	-0.068*	0.066***	-0.003	-0.001	-0.010
	(0.028)	(0.060)	(0.041)	(0.017)	(0.036)	(0.019)	(0.016)	(0.010)	(0.033)
Father incarcerated X exposure	0.037	0.079	-0.003	0.009	0.065	-0.049**	0.034*	0.012	0.063
	(0.035)	(0.065)	(0.042)	(0.019)	(0.045)	(0.024)	(0.019)	(0.010)	(0.040)
10th percentile, mothers	-.172***	-.203***	-.138**	-.095***	-.151***	-.024	-.022	-.007	-.042
	(.04)	(.069)	(.069)	(.031)	(.058)	(.027)	(.028)	(.017)	(.056)
50th percentile, mothers	-.16***	-.195***	-.119*	-.088***	-.139**	-.026	-.017	-.008	-.029
	(.039)	(.067)	(.064)	(.03)	(.056)	(.026)	(.027)	(.016)	(.055)
90th percentile, mothers	-.084	-.14**	.002	-.044	-.06	-.044	.018	-.012	.06
	(.052)	(.068)	(.066)	(.036)	(.054)	(.027)	(.028)	(.017)	(.069)
10th percentile, fathers	.017	-.011	.045	-.011	-.064*	.063***	-.002	0	-.007
	(.027)	(.058)	(.04)	(.017)	(.035)	(.019)	(.016)	(.01)	(.033)
50th percentile, fathers	.023	.003	.045	-.009	-.053	.055***	.004	.002	.004
	(.027)	(.054)	(.038)	(.017)	(.033)	(.017)	(.015)	(.009)	(.032)
90th percentile, fathers	.074	.111	.041	.003	.035	-.012	.05*	.018	.089
	(.056)	(.09)	(.062)	(.032)	(.067)	(.035)	(.029)	(.015)	(.064)
Dependent mean	0.241	0.297	0.175	0.083	0.126	0.032	0.043	0.013	0.078
Observations	155497	84117	71380	155460	84104	71356	131528	69312	62216

Incarceration instrumented by judge leave-out incarceration rate. All specifications include year X court fixed effects, and indicators for previous court appearances and incarcerations. Standard errors clustered at the judge level. * $p <$ 0.10, ** $p < 0.05$, *** $p < 0.01$.

Table 4.19. Effect of incarceration on juvenile criminal justice involvement

	All	Boys	Girls	All	Boys	Girls
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Appeared in court as a juvenile						
Parent incarcerated (=1)	-0.063***	-0.042	-0.074**			
	(0.022)	(0.039)	(0.035)			
Mother incarcerated (=1)				-0.071	-0.042	-0.094
				(0.060)	(0.098)	(0.067)
Father incarcerated (=1)				-0.057**	-0.042	-0.062
				(0.029)	(0.035)	(0.040)
Dependent mean	0.194	0.253	0.124	0.194	0.253	0.124
Observations	41733	22467	19266	41733	22467	19266
Panel B: Incarcerated as a Juvenile						
Parent incarcerated (=1)	-0.029**	-0.037	-0.015			
	(0.013)	(0.023)	(0.014)			
Mother incarcerated (=1)				-0.064**	-0.087*	-0.034
				(0.025)	(0.045)	(0.026)
Father incarcerated (=1)				-0.012	-0.013	-0.004
				(0.018)	(0.028)	(0.019)
Dependent mean	0.046	0.067	0.021	0.046	0.067	0.021
Observations	41733	22467	19266	41733	22467	19266

Incarceration instrumented by judge leave-out incarceration rate. All specifications include year X court fixed effects. Standard errors clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.20. Effect of incarceration on child juvenile court outcomes

	In court (=1)			Incarcerated (=1)		
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: All parents					
	All	Boys	Girls	All	Boys	Girls
Parent incarcerated (=1)	-0.157***	-0.081	-0.205***	-0.090**	-0.110*	-0.057**
	(0.044)	(0.068)	(0.054)	(0.037)	(0.060)	(0.028)
Incarcerated X exposure	0.181**	0.074	0.253***	0.115**	0.138	0.082
	(0.078)	(0.117)	(0.092)	(0.058)	(0.092)	(0.052)
10th percentile	-.148***	-.077	-.192***	-.084**	-.103*	-.053**
	(.041)	(.063)	(.05)	(.034)	(.056)	(.026)
50th percentile	-.119***	-.065	-.151***	-.065***	-.08*	-.04**
	(.03)	(.049)	(.041)	(.025)	(.042)	(.019)
90th percentile	.04	-.001	.072	.036	.041	.032
	(.051)	(.082)	(.07)	(.032)	(.05)	(.034)
Dependent mean	0.196	0.256	0.126	0.047	0.068	0.022
Observations	41848	22540	19308	41848	22540	19308
	Panel B: By gender of parent					
Mother incarcerated (=1)	-0.137*	-0.079	-0.190*	-0.140***	-0.182*	-0.079*
	(0.080)	(0.126)	(0.098)	(0.052)	(0.094)	(0.041)
Mother incarcerated X exposure	0.136	0.081	0.195	0.139*	0.168	0.089
	(0.145)	(0.197)	(0.153)	(0.083)	(0.141)	(0.084)
Father incarcerated (=1)	-0.178***	-0.093	-0.218***	-0.062	-0.071	-0.046
	(0.055)	(0.075)	(0.069)	(0.045)	(0.068)	(0.038)
Father incarcerated X exposure	0.222**	0.091	0.294**	0.099	0.116	0.078
	(0.108)	(0.141)	(0.147)	(0.078)	(0.117)	(0.078)
10th percentile, mothers	-.131*	-.075	-.181*	-.133***	-.174**	-.074**
	(.075)	(.12)	(.093)	(.049)	(.088)	(.038)
50th percentile, mothers	-.114*	-.065	-.157*	-.116***	-.153**	-.063**
	(.066)	(.108)	(.081)	(.04)	(.073)	(.031)
90th percentile, mothers	-.006	-.001	-.001	-.005	-.019	.008
	(.106)	(.147)	(.102)	(.044)	(.068)	(.055)
10th percentile, fathers	-.166***	-.088	-.202***	-.056	-.065	-.042
	(.05)	(.068)	(.063)	(.041)	(.062)	(.034)
50th percentile, fathers	-.128***	-.072	-.151***	-.039	-.045	-.028
	(.036)	(.049)	(.045)	(.03)	(.045)	(.023)
90th percentile, fathers	.174	.052	.248	.094	.113	.078
	(.127)	(.161)	(.179)	(.085)	(.127)	(.092)
Dependent mean	0.196	0.256	0.126	0.047	0.068	0.022
Observations	41848	22540	19308	41848	22540	19308

Incarceration instrumented by judge leave-out incarceration rate. All specifications include year X court fixed effects. Standard errors clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.21. School test scores and absenteeism on parental incarceration

	Year FE				Student FE			
	(1) Math	(2) Read	(3) PCA	(4) Absent	(5) Math	(6) Read	(7) PCA	(8) Absent
Parent incarcerated	0.134	-0.113	-0.0530	0.388				
	(0.286)	(0.176)	(0.200)	(4.208)				
Incarceration X post	-0.117	0.224	0.134	0.690	-0.275	-0.102	-0.198	-0.488
	(0.314)	(0.202)	(0.242)	(4.670)	(0.281)	(0.194)	(0.219)	(4.508)
Observations	60579	60385	60165	129080	56348	56041	53939	125445

Sample in header. Test scores and first principle component of math and reading test scores measured in standard deviations. Absences measured in days. Controls in include student race and age. Instruments include judge leave-out incarceration rate and interaction with post-assignment. Standard errors in parentheses and clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.22. Effect of incarceration on child outcomes

	In court (=1)			Incarcerated (=1)			Teen pregnancy (=1)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	By birth neighborhood SES quartiles								
	All	Boys	Girls	All	Boys	Girls	All	Boys	Girls
Incarcerated X Birth SES (Q1)	-0.133***	-0.193***	-0.063	-0.058***	-0.102**	-0.004	0.005	0.002	0.011
	(0.036)	(0.069)	(0.040)	(0.022)	(0.044)	(0.016)	(0.018)	(0.013)	(0.031)
Incarcerated X Birth SES (Q2)	-0.065*	-0.077	-0.064*	-0.054**	-0.090*	-0.019	-0.010	-0.008	-0.008
	(0.037)	(0.075)	(0.038)	(0.027)	(0.052)	(0.019)	(0.017)	(0.014)	(0.033)
Incarcerated X Birth SES (Q3)	0.041	0.037	0.027	0.003	-0.016	0.013	-0.005	-0.023	0.023
	(0.042)	(0.074)	(0.045)	(0.027)	(0.051)	(0.018)	(0.016)	(0.014)	(0.029)
Incarcerated X Birth SES (Q4)	0.032	-0.032	0.114**	-0.016	-0.052	0.028	-0.008	-0.014	-0.002
	(0.039)	(0.069)	(0.050)	(0.026)	(0.048)	(0.019)	(0.020)	(0.016)	(0.039)
Dependent mean	0.241	0.297	0.175	0.241	0.126	0.032	0.043	0.013	0.078
Observations	121212	65974	55238	121178	65962	55216	103651	55126	48525

Incarceration instrumented by judge leave-out incarceration rate. All specifications include year X court fixed effects, and indicators for previous court appearances, incarcerations, and socioeconomic status of birth address. Standard errors clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.23. Effect of incarceration on sibling outcomes

	Just mother	Just father	Mother	Father	Both	All
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: In court						
Sibling incarcerated (=1)	-0.072**	-0.047	-0.049	0.004	0.050	-0.050*
	(0.034)	(0.080)	(0.032)	(0.045)	(0.075)	(0.028)
Observations	64819	14792	83313	33286	18494	98105
Panel B: Incarcerated						
Sibling incarcerated (=1)	-0.073***	-0.008	-0.057***	0.005	0.010	-0.048***
	(0.023)	(0.037)	(0.019)	(0.030)	(0.048)	(0.017)
Observations	64801	14788	83295	33282	18494	98083

All specifications include year X court fixed effects, and indicators for previous court appearances and incarcerations. Standard errors clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.24. Effect of incarceration on sibling outcomes

	Just mother	Just father	Mother	Father	Both	All
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: In court						
Incarcerated X Birth SES (Q1)	-0.241***	-0.204*	-0.217***	-0.128	-0.112	-0.241***
	(0.065)	(0.109)	(0.063)	(0.086)	(0.124)	(0.065)
Incarcerated X Birth SES (Q2)	-0.166***	-0.049	-0.171***	-0.103	-0.204**	-0.166***
	(0.060)	(0.079)	(0.057)	(0.064)	(0.102)	(0.060)
Incarcerated X Birth SES (Q3)	0.119*	0.028	0.118**	0.082	0.116	0.119*
	(0.068)	(0.128)	(0.059)	(0.075)	(0.100)	(0.068)
Incarcerated X Birth SES (Q4)	0.054	0.027	0.086	0.113	0.189*	0.054
	(0.055)	(0.131)	(0.054)	(0.083)	(0.113)	(0.055)
Dependent mean	0.243	0.146	0.238	0.189	0.223	0.225
Observations	38205	8978	48781	19553	10567	38205
Panel B: Incarcerated						
Incarcerated X Birth SES (Q1)	-0.130***	0.006	-0.121***	-0.024	-0.085	-0.130***
	(0.043)	(0.055)	(0.038)	(0.050)	(0.083)	(0.043)
Incarcerated X Birth SES (Q2)	-0.105**	0.009	-0.108**	-0.041	-0.124	-0.105**
	(0.045)	(0.050)	(0.042)	(0.052)	(0.086)	(0.045)
Incarcerated X Birth SES (Q3)	-0.049	0.058	-0.048	0.026	-0.013	-0.049
	(0.040)	(0.063)	(0.034)	(0.056)	(0.083)	(0.040)
Incarcerated X Birth SES (Q4)	-0.076**	0.004	-0.053	0.018	0.013	-0.076**
	(0.038)	(0.058)	(0.037)	(0.052)	(0.082)	(0.038)
Dependent mean	0.089	0.041	0.086	0.061	0.076	0.079
Observations	38199	8977	48775	19552	10567	38199

Incarceration isntrumented by judge leave-out incarceration rate. All specifications include year X court fixed effects, and indicators for previous court appearances and incarcerations. Standard errors clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Bibliography

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

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

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

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

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

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

Arendt, J. (2000). Melatonin, circadian rhythms, and sleep. *New England Journal of Medicine*, 343(15):1114–1116.

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

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

Campbell, I. G., Grimm, K. J., de Bie, E., and Feinberg, I. (2012). Sex, puberty, and the timing of sleep EEG measured adolescent brain maturation. *Proceedings of the National Academy of Sciences*, 109(15):5740–5743.

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

Carrell, S. E., Maghakian, T., and West, J. E. (2011). A's from ZZZZ's? The causal effect of school start time on the academic achievement of adolescents. *American Economic Journal: Economic Policy*, 3(3):62–81.

Carskadon, M. A., Acebo, C., and Jenni, O. G. (2004). Regulation of adolescent sleep: Implications for behavior. *Annals of the New York Academy of Sciences*, 1021(1):276–291.

Carskadon, M. A., Acebo, C., Richardson, G. S., Tate, B. A., and Seifer, R. (1997). An approach to studying circadian rhythms of adolescent humans. *Journal of biological rhythms*, 12(3):278–289.

Carskadon, M. A., Vieira, C., and Acebo, C. (1993). Association between puberty and delayed phase preference. *Sleep*, 16(3):258–258.

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

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

Chen, X., Heckman, J. J., Vytlacil, E., et al. (2000). Identification and sqrt n efficient estimation of semiparametric panel data models with binary dependent variables and a latent factor. In *Econometric Society World Congress 2000 Contributed Papers*, number 1567. Econometric Society.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. Technical Report 9.

Crowley, S. J., Acebo, C., and Carskadon, M. A. (2007). Sleep, circadian rhythms, and delayed phase in adolescence. *Sleep Medicine*, 8(6):602–612.

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

Dahl, G. B. and Lochner, L. (2012). The impact of family income on child achievement: Evidence from the earned income tax credit. *The American Economic Review*, 102(5):1927–1956.

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

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

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

Douglas, M. (1999). Late to bed, early to rise makes a teen-ager ... tired. *New York Times*.

Doyle, J. J. (2008). Child protection and adult crime: Using investigator assignment to estimate causal effects of foster care. *Journal of political Economy*, 116(4):746–770.

Eaton, D. K., McKnight-Eily, L. R., Lowry, R., Perry, G. S., Presley-Cantrell, L., and Croft, J. B. (2010). Prevalence of insufficient, borderline, and optimal hours of sleep among high school students–united states, 2007. *Journal of Adolescent Health*, 46(4):399–401.

Edwards, F. (2012). Early to rise? the effect of daily start times on academic performance. *Economics of Education Review*, 31(6):970–983.

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

Eren, O. and Mocan, N. (2016). Emotional judges and unlucky juveniles. Technical report, National Bureau of Economic Research.

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

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

Fogel, S. M. and Smith, C. T. (2011). The function of the sleep spindle: a physiological index of intelligence and a mechanism for sleep-dependent memory consolidation. *Neuroscience & Biobehavioral Reviews*, 35(5):1154–1165.

Frakes, M. D. and Wasserman, M. F. (2014). Is the time allocated to review patent applications inducing examiners to grant invalid patents?: Evidence from micro-level application data. *Review of Economics and Statistics*, (0).

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

Gibson, M. and Shrader, J. (2015). Time use and productivity: The wage returns to sleep. *Working paper, University of California San Diego*.

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

Grant, A. G. and Rehaag, S. (2015). Unappealing: An assessment of the limits on appeal rights in canada's new refugee determination system.

Groen, J. A. and Pabilonia, S. W. (2015). Snooze or lose: High school start times and academic achievement. Technical report, Bureau of Labor Statistics, US Department of Labor.

Group, A. S. W. and Committee on Adolescence, A. A. o. P. (2014). School start times for adolescents. *Pediatrics*.

Hansen, M., Janssen, I., Schiff, A., Zee, P. C., and Dubocovich, M. L. (2005). The impact of school daily schedule on adolescent sleep. *Pediatrics*, 115(6):1555–1561.

Hanushek, E. A. (1974). Efficient estimators for regressing regression coefficients. *The American Statistician*, 28(2):66–67.

Hausegger, L., Riddell, T., Hennigar, M., and Richez, E. (2010). Exploring the links between party and appointment: Canadian federal judicial appointments from 1989 to 2003. *Canadian Journal of Political Science/Revue canadienne de science politique*, 43(3):633–659.

Hinrichs, P. (2011). When the bell tolls: The effects of school starting times on academic achievement. *Education*, 6(4):486–507.

Jacob, B. A. and Rockoff, J. E. (2011). Organizing schools to improve student achievement: Start times, grade configurations, and teacher assignments. *The Hamilton Project*, pages 1–28.

Jenni, O. G. and Carskadon, M. A. (2012). Sleep behavior and sleep regulation from infancy through adolescence: Normative aspects. *Sleep Medicine Clinics*, 7(3):529–538.

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

Krueger, A. B. and Whitmore, D. M. (2002). *Bridging the Achievement Gap*, chapter Would smaller classes help close the black-white achievement gap? Brookings Institution Press.

Laberge, L., Petit, D., Simard, C., Vitaro, F., Tremblay, R., and Montplaisir, J. (2001). Development of sleep patterns in early adolescence. *Journal of Sleep Research*, 10(1):59–67.

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

Lufi, D., Tzischinsky, O., and Hadar, S. (2011). Delaying school starting time by one hour: Some effects on attention levels in adolescents. *Journal of Clinical Sleep Medicine: Official Publication of the American Academy of Sleep Medicine*, 7(2):137.

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

Maquet, P., Laureys, S., Peigneux, P., Fuchs, S., Petiau, C., Phillips, C., Aerts, J., Del Fiore, G., Degueldre, C., Meulemans, T., et al. (2000). Experience-dependent changes in cerebral activation during human rem sleep. *Nature neuroscience*, 3(8):831–836.

Marshall, W. A. and Tanner, J. M. (1970). Variations in the pattern of pubertal changes in boys. *Archives of disease in childhood*, 45(239).

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

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

NCES (2012). Average start time for public high schools and percentage distribution of start times in public high schools, by selected school characteristics. *Schools and Staffing Survey, National Center for Education Statistics*.

Ng, E., Ng, D., and Chan, C. (2009). Sleep duration, wake/sleep symptoms, and academic performance in Hong Kong secondary school children. *Sleep and Breathing*, 13(4):357–367.

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

Philippe, A. and Ouss, A. (2016). "no hatred or malice, fear or affection": Media and sentencing.

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

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

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

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458.

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

Sadeh, A., Gruber, R., and Raviv, A. (2003). The effects of sleep restriction and extension on school-age children: What a difference an hour makes. *Child development*, 74(2):444–455.

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

Schanzenbach, D. W. (2006). What have researchers learned from Project STAR? *Brookings papers on education policy*, 9:205–228.

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

Smith, A. C. (2016). Spring forward at your own risk: Daylight saving time and fatal vehicle crashes. *American Economic Journal: Applied Economics*, 8(2):65–91.

Wahlstrom, K., Wrobel, G., Kubow, P., et al. (1998). Minneapolis Public Schools start time study executive summary 1998.

Walker, M. P. and Stickgold, R. (2006). Sleep, memory, and plasticity. *Annual Review of Psycholoft*, 57(1):139–166.

APPENDIX A

A.1. Online Appendix for *Rise and Shine: The Effect of School Start Times on Academic Performance from Childhood through Puberty*

A.1.1. Robustness checks for mover definition

Our identifying variation comes from students who move between schools in different time zones in the Florida panhandle. Most of these moves are quite long-distance; the median move is 83 miles. The disruption inherent in such a move may have an independent effect on achievement, which is important to control for in our context. To help identify the effect of moving, as well as the effect of other school-level covariates, we include in our sample students who move within a time zone. This requires defining what constitutes a move by setting a threshold distance between the schools the student attended. Otherwise, graduating from middle school to high school would constitute a move. A high threshold has the advantage of making the move more likely to match a cross-time zone move in terms of disruptiveness; a low threshold increases sample size and precision.

We settled on a threshold of 25 miles, but our results are robust to other threshold choices. Table A1 presents estimates for 15, 20, 25, and 30 mile thresholds for math and reading outcomes. We also consider defining a move as any move between different school districts, although this will include students who move less disruptive distances, such as when families move to a nearby suburb that happens to be in a different district. Across all definitions, the

results are broadly consistent. In math, the effect for prepubescent children ranges from 0.009 to 0.037 SDs; the effect for adolescents ranges from 0.067 to 0.084 SDs. In reading, the range is 0.034 to 0.061 for younger children and 0.044 to 0.057 for adolescents. The effects statistically differ from zero for adolescents for both math and reading across all distances.

A.1.2. Specification robustness checks

We include two sets of control variable robustness checks. First, in Table A2, we consider different levels of aggregation for the demographic share controls (FRL, male, black, Asian, and Hispanic). Instead of aggregating at the school-year level, as we do in our main results, we consider district-year, district third graders-year,[1] school-year, and school-grade-year. All specifications include age-gender dummies and an individual fixed effect. For each level of aggregation, we present one specification with no other controls, one that adds urban dummies and log income controls, and a final model that includes school size and student/teacher ratio.

Comparing across the rows of Table A2, the results are largely unchanged. In Panel A, all specifications show an effect size in math of 0.003-0.037 SDs for prepubescents, and 0.062-0.096 for adolescents. The effect is statistically significant at the the 1% level or better for adolescents but null for younger students. In reading, the estimates are also similar across specifications: 0.046-0.087 SDs for prepubescents, and 0.044-0.074 SDs for adolescents. The prepubescent effect is occasionally significant at the 5% level; the adolescent effect has a p-value of about 1%.

For absences, the inclusion of demographics (but not the level of aggregation) makes a substantive difference in the results. Comparing Columns 1-3 with Columns 4-15, the inclusion

[1]District third graders-year is the demographic means for the third graders in the given district-year.

of demographic controls (at any level of aggregation) reduces the size of the suspension effect from about 1.5 percentage points and significant at the 1% level to about 0.8 percentage points and significant at the 10% level for prepubescents. The adolescent effects are generally null once we control for demographics. Since there may be significant between-school differences in policies for counting absences (and these may be correlated with school demographics), we think that the results with demographic controls are more trustworthy. It is therefore reassuring that they are the same regardless of the level of demographic aggregation.

Table A3 contains sur second control robustness check. Columns 1 and 3 restate our baseline results for math and reading. Columns 2 and 5 include controls for latitude; average sunrise times over the school year vary by about a minute over the north-south range of the panhandle[2] and this could conceivably have some affect on sleep (in contrast, the east-west variation in sunrise times from longitude is nearly 20 minutes, excluding the time zone change). The addition of latitude has a moderately sized but statistically insignificant effect on the prepubescent coefficients. The change in the adolescent coefficients is smaller.

In Columns 3 and 6 of Table A3 we test whether the inclusion of third grade district test scores as control variables affects the results. Third grade test scores are appealing as a summary measure of district quality, but may be endogeneous if start times affect performance for children in kindergarten to third grade. For this reason we do not include them in our main specification, but it is reassuring that they have little effect on the results.

[2]The average disguises some larger differences over the year; but it is never larger than three minutes.

A.1.3. Changes in school characteristics over the move

A potential threat to our identification strategy is changes in school and peer characteristics as students move between time zones. If students moving from CT to ET move to significantly worse schools, while ET-CT movers moved to better schools, it would not be surprising that student achievement declined upon entering ET and rose upon exiting. Because, on average, there is less sunlight before school in ET than in CT, this could generate a spurious positive relationship between relative school start times and academic achievement.

We consider this question directly in Table A4. We take the years directly before and after each move, and term these pairs of years a *moving episode*.[3] We then regress school- and zip-level characteristics on moving episode fixed effects and move indicators for the four different types of movers: Eastern-Eastern, Central-Central, Eastern-Central, and Central-Eastern. Each coefficient is a measure of the change in characteristics over the move. As outcomes, we consider the five school-level demographic share controls included in our preferred specification (percent FRL, male, black, Asian, and Hispanic), as well as school student/teacher ratio and zipcode-level median income as a measure of school and community resources.

The first two rows of Table A4 show that peer quality changed slightly over the move for within-time zone movers. ET-ET movers had 4.5 percentage points fewer FRL classmates; CT-CT movers had 1.7 percentage points fewer. School quality as measured by the student/teacher ratio increased slightly for both groups. Median income rose by $1,000 for within-ET movers and fell by $430 for within-CT movers. These differences are statistically significant, but none are particularly large or striking.

[3]Since occasionally a student will move in consecutive years, a small number of observations are repeated.

The cross-time zone movers tell a slightly different story. Eastward movers generally ended up in a richer area — 4.5 percentage points fewer FRL classmates and $5,700 higher median income — and had 14.0 percentage points more black classmates and 0.5 percentage points more Hispanic classmates. School quality as measured by the student/teacher ratio was unchanged. ET-CT movers saw approximately the opposite changes in medium income and percent of black students. The economic and peer changes may work in opposite directions in this case, making it unclear in which direction the overall bias goes. However, neither the inclusion of demographic controls (in Table 4.2) or income controls (in Table A2) substantively changes our results, suggesting that changes in peer characteristics have only a moderate effect on outcomes over the move, and do not significantly affect our results.

A.1.4. Performance trend before move

In Section 1.4.1, we show that test score trends are similar for all groups of movers in the years before the move. However, math scores trend *up*, which is somewhat surprising since the disruption of the upcoming move would be expected to reduce scores. Figure A1 show results from a regression of scale scores on time-until-move dummies and a fixed effect for the period until the move. This is identical to the regression displayed in Figure 4.1, but without controls. The Figure confirms that unconditionally, test scores trend down in both math and reading before a move. This is largely a result of removing the age-gender fixed effects, which soak up any time trend. Comparing across different groups of movers, the trends are slightly further apart than in the version with controls, but are still generally statistically indistinguishable.

A.1.5. Robustness checks for puberty definition

One of our main interests in this paper is how the effect of relative school start times varies with pubertal status. This requires a working definition of puberty, and there are several defensible alternatives. Pubertal development is typically measured with the Tanner Scale. There are two versions; one that uses levels of pubic hair to define the stages and another that uses breast and genital development. We rely on the pubic hair version of the Scale, which Campbell et al. (2012) indicate is more closely associated with pubertal changes in sleep patterns. They also note that changes in sleep patterns begin during Stage 3, so we use the age of median attainment (by gender) of Stage 3 as the definition of puberty.

Table A5 shows our main results with three alternative definitions of puberty: pubic hair Stage 2, pubic hair Stage 4, and breast/genital Stage 3. These changes typically shift the age of puberty by at most a year, and not necessarily for both genders. The results are largely unchanged, although slightly attenuated in some specifications. Because this definition of puberty is a worse fit for the underlying biological processes, this is unsurprising.

A.1.6. Estimates without interactions

Table A6 displays a version of our baseline model without an interaction between relative start time and pubertal status. Allowing for heterogeneity by pubertal status is important, but for completeness we have included this specification.

Across the rows, the change in sunlight is about 30 minutes over the time zone border. For both math and reading, the effect of moving start times one hour later is about the average of the child and adolescent effects from Table 4.2. In math, the estimated effect is 0.043 SD per hour by the final column, and the estimates are only occasionally statistically significant. In reading,

the effect is 0.059 SD per hour by the final column, and the effect sizes are all significant to at least the 5% level in all estimates. The attendance results vary, with a decrease of 0.7 percentage points in absence per hour of sunlight by the final column.

A.1.7. PSID data definitions

In this paper, we demonstrate that students treated with later relative start times have higher academic achievement. However, we do not directly observe sleep levels in the academic outcomes dataset. To more concretely link changes in start times to changes in sleep, we use the Child Development Supplement of the Panel Study of Income Dynamics (PSID) to estimate the effect of the time zone boundary on sleep. The survey collected time use diaries for students on a weekend day and a weekday in the years 1997, 2002, and 2007. We include all states with a single time zone,[4] and all children who were 6-19 during the survey and within 400 miles of the ET-CT time zone boundary. Our aim is descriptive, so we regress daily hours of sleep on a fully interacted set of dummies for puberty, CT, and whether the night was a weekend. In our preferred specification, we also include controls for gender, black/non-black, and FRL status. We expect that children in CT will have more sleep on weekdays when they face earlier relative start times, and those in ET will compensate with more sleep on weekends.

Table A7 contains the results. As discussed in Section 1.5.3, children in CT get 6 minutes more sleep per night during the week than children in ET; during puberty they get 17 minutes more. On the weekend, children in ET compensate for low levels of sleep during the week by sleeping 10 minutes more per night in the years before puberty and 19 minutes more while in puberty. We conservatively cluster at the state level. The coefficient for the difference in sleep

[4]The CDS does not geocode individuals at a sub-state level in the publicly available version, which precludes analysis using observations in states with multiple time zones — including Florida.

between adolescents in CT and ET is significant at the 10% level; most others are not. Including student fixed effects suggests a slightly larger difference between the time zones: the decrease in sleep during puberty is 15 minutes smaller for adolescents in CT than in ET. This set of results corresponds to a pass-through rate of about 40-50% from school start times to sleep if Florida panhandle school start times are representative of the rest of the US near the ET-CT time zone boundary. This number is close to the 46% pass-through reported by Wahlstrom (1998).

A.1.8. Treatment bleed for schools near the time zone boundary

In the placebo analysis, we study how test scores change when students move east-west or west-east but *not* across the true time zone boundary. Ideally, we would examine within-time zone moves to and from the region directly adjacent to the boundary, to help test whether there are unobservable changes in the school or community environment that occur nearby, but not exactly at, the time zone boundary.

This approach will be problematic if there is an effect of being near the time zone boundary on school start times —- then, moving from directly beside the boundary in CT to a city fifty miles west could increase relative start times, directly increasing test scores. Figure A4 displays a nonparametric regression of relative start times on distance to the time zone boundary, estimated separately for each time zone. In the region directly adjacent to the boundary, start times veer towards the other time zone's norm, particularly for adolescents. We interpret this as the synchronization of start times across time zones, which allows parents to help their children prepare for school before going to work, whether or not they are commuting across time zones. This also means that start times are later for students moving west either from the region directly beside the boundary in CT, or *to* the region directly beside the boundary in ET.

In the main placebo results, we account for the treatment bleed across time zones by taking out a 25 mile "donut" around the time zone boundary. However, in the interest of completeness we include the unexcised version in Figure A5. The difference with Figure 4.8 is most stark in the puberty-time zone coefficient for math, where there is a consistent effect above the size of the true coefficient. Comparing between figures, removing the donut around the time zone boundary reduces the size of *all* placebo coefficients. The placebo effect is coming largely from individuals moving between the area close to the true time zone boundary and the rest of the study area, not individuals moving between areas far from the time zone boundary.

A.2. Appendix for *Judicial Errors: Evidence from Refugee Appeals*

A.2.1. Details on the use of regressors for identification

With a small change of notation, the main model of Section 2.2 can be recast as a single-spell duration model (**?**), where the "duration" is the amount of time until a judge rejects the applicant's case (duration is capped at 2). Equation 2.1 from the main text sets out the problem as identifying the parameters of the choice model where approval in each stage s occurs if

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

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

(1) r_i and $\widetilde{\varepsilon}_{ijs}(W_{ijs})$ are independent and have a median of zero. The variance of r_i is known and finite, and the variances of $\widetilde{\varepsilon}_{ijs}(W_{ijs})$ are finite.

(2) $X_{ij1}\beta_1 | \gamma_j, W_{ij1}$ is continuous with large support.

(3) $X_{ik2}\beta_2 | X_{ik1}\beta_1, \gamma_j, \gamma_k, W_{ij1}, W_{ij2}$ is continuous with large support.

(4) At least one component of β_1 is assumed equal to the same component in β_2.

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

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

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

where $H_{js,W}$ is the distribution of $\eta_{ks} = \widetilde{\varepsilon}_{ijs}(W_{ijs}) - r_i$, the composite error of the refugee-level equality variable r_i and the case-judge idiosyncratic error $\widetilde{\varepsilon}_{ijs}(W_{ijs})$. As in Manski (1975), the assumption of median-zero errors allows nonparametric identification of β_1 and H_{k1} up to scale. However, the identity of judge j and the regressors W_{ijs} enter the distribution $H_{js,W}$, and thus neither W_{ijs} or the judge effect γ_j can be used for identification. Instead, X_{ij1} traces out the distribution of $H_{js,W}$, which is why Assumption 2 calls for large support conditional on judge assignment and W_{ijs}.

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

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

$$= H_{k2,W}(-X_{ij2}\beta_2 - \gamma_{j2})$$

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

A.2.2. Estimation details

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

A.2.3. First round approval

$$\begin{aligned} P(r-\varepsilon_1>0) &= \int_0^{\infty} P(r>\tilde{\varepsilon}_1)dF(\tilde{\varepsilon}_1) \\ &= \int_0^{x_m} P(r>\tilde{\varepsilon}_1)dF(\tilde{\varepsilon}_1)+\int_{x_m}^{\infty} P(r>\tilde{\varepsilon}_1)dF(\tilde{\varepsilon}_1) \end{aligned} \tag{A.3}$$

The first term in Equation A.3 can be shown to be equal to $\Phi[\frac{\ln(x_m)-\mu_1}{\sigma_1}]$. Then,

$$\begin{aligned} \int_{x_m}^{\infty} P(r>\tilde{\varepsilon}_1)dF(\tilde{\varepsilon}_1) &= \int_{x_m}^{\infty} \frac{x_m^{\alpha}}{\tilde{\varepsilon}_1^{\alpha}}\phi\left(\frac{\ln(\tilde{\varepsilon}_1)-\mu_1}{\sigma_1}\right)\frac{1}{\sigma_1\tilde{\varepsilon}_1}d\tilde{\varepsilon}_1 \\ &= x_m^{\alpha}\int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}}^{\infty} e^{-\alpha(\sigma_1 y+\mu_1)}\phi(y)dy \\ &= x_m^{\alpha}e^{-\alpha\mu_1}\int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}}^{\infty} e^{-\alpha\sigma_1 y-\frac{y^2}{2}}dy \\ &= x_m^{\alpha}e^{-\alpha\mu_1+\frac{\alpha^2\sigma_1^2}{2}}\int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}}^{\infty} e^{-\frac{1}{2}(y+\alpha\sigma_1)^2}dy \\ &= x_m^{\alpha}e^{-\alpha\mu_1+\frac{\alpha^2\sigma_1^2}{2}}\int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}+\alpha\sigma_1}^{\infty} e^{-\frac{1}{2}y^2}dy \\ &= x_m^{\alpha}e^{-\alpha\mu_1+\frac{\alpha^2\sigma_1^2}{2}}\left[1-\Phi\left(\frac{\ln(x_m)-\mu_1}{\sigma_1}+\alpha\sigma_1\right)\right] \end{aligned} \tag{A.4}$$

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

A.2.3.1. Approval in both rounds. In the model I estimate, occasionally the same judge is assigned to make the first and second round decision for a defendant. I model this by allowing between-round errors to be correlated whenever it is the same judge and estimate the correlation as an additional parameter. Below, I present the full derivations for the no-correlation case (which is more intuitive), then explain how the model works with correlations.

The likelihood of approval in the first round is

$$P(r > \varepsilon_2 \cap r > \varepsilon_1) = \int_0^\infty \int_0^\infty P(r > \tilde{\varepsilon}_2 | r > \tilde{\varepsilon}_1) P(r > \tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_2) \tag{A.5}$$

The terms inside the integrals can be rewritten

$$P(r > \tilde{\varepsilon}_1) = \mathbb{1}[\tilde{\varepsilon}_1 < x_m] + \mathbb{1}[\tilde{\varepsilon}_1 \geq x_m] \frac{x_m^\alpha}{\tilde{\varepsilon}_1^\alpha} \tag{A.6}$$

and

$$\begin{aligned} P(r > \tilde{\varepsilon}_2 | r > \tilde{\varepsilon}_1) = & \mathbb{1}[\tilde{\varepsilon}_1 < x_m] \left[\mathbb{1}[\tilde{\varepsilon}_2 < x_m] + \mathbb{1}[\tilde{\varepsilon}_2 \geq x_m] \frac{x_m^\alpha}{\tilde{\varepsilon}_2^\alpha} \right] + \\ & \mathbb{1}[\tilde{\varepsilon}_1 \geq x_m] \left[\mathbb{1}[\tilde{\varepsilon}_2 < \tilde{\varepsilon}_1] + \mathbb{1}[\tilde{\varepsilon}_2 \geq \tilde{\varepsilon}_1] \frac{\tilde{\varepsilon}_1^\alpha}{\tilde{\varepsilon}_2^\alpha} \right] \end{aligned} \tag{A.7}$$

Substituting into Equation A.5 and expanding the integrals,

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

Further separate the integrals into four components:

(A.8) $$\int_0^{x_m} \int_0^{x_m} dF(\tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_2)$$

(A.9) $$\int_{x_m}^{\infty} \int_0^{x_m} \frac{x_m^{\alpha}}{\tilde{\varepsilon}_2^{\alpha}} dF(\tilde{\varepsilon}_1) dF(\tilde{\varepsilon}_2)$$

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

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

These four equations (A.8-A.11) are all simple to evaluate because the distribution of a Pareto-distributed random variable conditional on being larger than a given threshold is itself Pareto. I solve them in turn:

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

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

The last two make use of the additional fact that

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

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

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

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

$$=x_m^{\alpha}e^{-\alpha\mu_1}\int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}}^{\infty}e^{-\alpha\sigma_1 y}\Phi\left(\frac{\sigma_1 y+\mu_1-\mu_2}{\sigma_2}\right)\phi(y)\,dy$$

$$=x_m^{\alpha}e^{-\alpha\mu_1+\frac{\alpha^2\sigma_1^2}{2}}\int_{\frac{\ln(x_m)-\mu_1}{\sigma_1}+\alpha\sigma_1}^{\infty}\Phi\left(\frac{\sigma_1 y-\alpha\sigma_1^2+\mu_1-\mu_2}{\sigma_2}\right)\phi(y)\,dy$$

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

$$=x_m^{\alpha}e^{-\alpha\mu_1+\frac{\alpha^2\sigma_1^2}{2}}BvN(\frac{(\mu_1-\mu_2)/\sigma_1-\alpha\sigma_1}{\sqrt{\sigma_2^2/\sigma_1^2+1}},-\frac{\ln(x_m)-\mu_1}{\sigma_1}-\alpha\sigma_1,$$

$$\frac{1}{\sqrt{\sigma_2^2/\sigma_1^2+1}},)$$

$$\int_{x_m}^{\infty}\int_{\tilde{\varepsilon}_1}^{\infty}\frac{x_m^{\alpha}}{\tilde{\varepsilon}_2^{\alpha}}dF_1dF_2=\int_{x_m}^{\infty}x_m^{\alpha}e^{-\alpha\mu_2+\frac{\alpha^2\sigma_2^2}{2}}\left[1-\Phi\left(\frac{\ln(\tilde{\varepsilon}_1)-\mu_2}{\sigma_2}+\alpha\sigma_1\right)\right]dF(\tilde{\varepsilon}_1)$$

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

$$\phi(y)dy\}$$

$$=\tilde{B}\{\left[1-\Phi\left(\frac{\ln(x_m)-\mu_1}{\sigma_1}\right)\right]-BvN(\frac{(\mu_1-\mu_2+\alpha\sigma_2^2)/\sigma_1}{\sqrt{\sigma_2^2/\sigma_1^2+1}},-\frac{\ln(x_m)-\mu_1}{\sigma_1},$$

$$\frac{1}{\sqrt{\sigma_2^2/\sigma_1^2+1}})\}$$

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

A.2.3.2. Approval in both rounds with error correlation. In this section I describe how the probabilities can be modified to allow for correlation between rounds. This is used when the same judge sees the case in both rounds. I describe the version for Equation A.11 in detail; the same method works for all the joint first- and second-round probabilities.

$$\int_{x_m}^{\infty}\int_{\tilde{\varepsilon}_1}^{\infty} \frac{x_m^{\alpha}}{\tilde{\varepsilon}_2^{\alpha}} dF_2 dF_1$$

$$= \int_{x_m}^{\infty}\int_{\tilde{\varepsilon}_1}^{\infty} \frac{x_m^{\alpha}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}\varepsilon_1\varepsilon_2} e^{-\alpha\ln(\varepsilon_2) - \frac{1}{2(1-\rho^2)}\left((\frac{\ln(\varepsilon_1)-\mu_1}{\sigma_1})^2 - 2\rho(\frac{\ln(\varepsilon_1)-\mu_1}{\sigma_1})(\frac{\ln(\varepsilon_2)-\mu_2}{\sigma_2}) + (\frac{\ln(\varepsilon_2)-\mu_2}{\sigma_2})^2\right)}$$

$$d\varepsilon_2 d\varepsilon_1$$

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

Complete the square in the exponentiated part, then substitute into the above equation. This allows you to take the integral with respect to x, leaving

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

Rearrange the term in the normal:

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

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

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

$$\phi(y)dy$$

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

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

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

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

A.2.4. MTE of first-round approval on second-round approval

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

$$\Delta^{MTE}(u_D) = E[Y_1 - Y_0 | U_D = u_d] \tag{A.12}$$

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

$$D^* = P(Z) \geq U_D \tag{A.13}$$

where U_D is normalized to be unit uniform and $P(Z)$ is the probability of treatment given assignment to the instrument Z. I use the first-round judge assignment as the instrument Z. As seen in Figure A7, the support of the instrument ranges from 0.03 to 0.28, with a large gap before a point mass at 0.70. I estimate the MTE using all observations, and using only the main mass. Nonetheless, the results are only identified by functional form for points larger than 0.3, and should be treated with caution.

As Figure A7 shows, there is not very much variation in MTE over the range of first-round judges; from the 3rd to the 28th percentile of the distribution of refugee quality, the approval probability drops from 46% to 35%, though this result is somewhat sensitive to the outlier judge. In other words, most judges' marginal rejections would have at least a one-third chance of approval in the second round.

A.2.5. Survey questions

As I discuss in Section 2.4.8, I fielded a survey of lawyers who had appeared in front of the Federal Court justices in my sample. The goal of the survey was to generate expert measures of the same parameters that are identified by my structural model.

From the court records, I located the names of 931 lawyers who had appeared in front of one of the judges in my sample. I was able to find online contact information for 551 of them.[5] In April 2017, I contacted the lawyers and requested that they fill out an online survey on their experience with Federal Court judges. After one reminder email, 64 lawyers responded for an overall response rate of 14%.[6] Table A11 compares responders to non-responders and lawyers for whom I couldn't find contact information. The main differences are that responders are more successful, with a first-round approval rate of 27% versus 19% for non-responders (the contacted sample is mostly lawyers for the claimants; government lawyers were included in the sample but their names are recorded much less frequently in the court documents). Respondents are slightly younger, with their first recorded case coming about one year later.

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

---

[5]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.

[6]This response rate compares favorably to telephone political polls, where response rates are below 10% (**?**). However, it is significantly lower than the 20% response rate for an email poll conducted by **?** 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).

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

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

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

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

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

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

I expected that the first question would be related to the judge-specific threshold γ, and the second question with the variance of the observational error σ_j. By design, the second question encompasses the two distinct aspects of σ_j detailed in Section 2.2.3.1. First, asking about predictability concerns test-retest reliability — will the judge understand the merits of the case? On the other hand, asking whether the judge decides cases on similar grounds as other judges is trying to unearth information about how judges consistently value different aspects of the case (inter-rater reliability), such as the relative weight they place on procedural versus substantive merits.

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

The main results are in Table 4.11, where I include only the first two questions. I discuss these in Section 2.4.8. The final question of the survey, which asked about how accurate the judge is, I did not discuss in the main text. This question does not have as clear an interpretation as the other two. There is no direct mapping of accuracy into the model, since accuracy implies a normative judgement about the correct outcome of the case. Reported accuracy is correlated positively with favorability and negatively with inconsistency, but more strongly with the former ($\rho = 0.7$ versus -0.46). Anecdotally, many of the lawyers that I corresponded with about the survey were involved in refugee-rights non-profits, so it is likely that they believe the claimants should win more cases than they currently do. Table A12 adds accuracy to the regression of model coefficients on survey responses; with no other regressors higher accuracy predicts lower

second-stage thresholds γ_{j2}, but this disappears when favorability and inconsistency are added. The relationship between favorability and γ_2, and inconsistency and σ_1 is almost unchanged.

A.2.6. Ramifications for judge-assignment IVs

Exploiting random judge assignment is an increasingly popular identification strategy (Aizer and Doyle, 2013; Dahl et al., 2013). The monotonicity condition in this context is simple: it requires that for any two judges, any individual convicted by the more lenient judge must also be convicted by the less lenient judge.[7] Mueller-Smith (2014) discusses how this can be violated when the researcher does not separately estimate judge severity by type of crime. If a judge is harsh for (say) drug crimes but lenient for violent crimes, on average they would be considered a medium-severity judge. But exposure to this judge versus one that uniformly sentences defendants for an average sentence would be a negative shock for a drug-crime defendant and a positive shock for a violent-crime defendant, generating defiers. Mueller-Smith shows how this problem can be circumvented when the econometrician has access to a rich set of covariates. The strategy is to use a LASSO first stage to select the instruments (in his case, judge and prosecutor effects interacted with defendant characteristics and crime type) with the best predictive power, ensuring that inconsistency associated with the observables is not contributing to violations of monotonicity. Other researchers have approximated this approach with simple interactions between judge assignment and crime type (**?**). Bhuller et al. (2016) show that their judge-assignment instrument (judge-mean incarceration rate on all cases) also predicts incarceration for subsets of the sample defined by defendant and charge characteristics. In most situations, these methods are likely to assuage first-order concerns about monotonicity violations.

[7] **?** shows that even under violations of monotonicity it may be possible to identify a less-interpretable LATE for a subset of compliers.

As I show in the main text, in my context there are high levels of inconsistency (and corresponding violations of monotonicity). This is not necessarily an indictment of previous research that relies on examiner-assignment instruments. As I discuss above, it is possible to partially test for monotonicity and construct instruments that are robust against most sorts of violations. Furthermore, Federal Court refugee appeals are very different than criminal courts or the SSDI system, and it is possible that these decisions are more susceptible to inconsistency for two reasons. First, appeal cases are almost by definition more marginal. About 60% of initially-rejected claimants appeal to the Federal Court, meaning that the cases on the docket are relatively difficult. Second, as I discuss in Section 2.3.2, there is no written precedent for first stage decisions. This makes it hard for judges to learn how other judges have acted in a similar situation. I show suggestive evidence in the main text that this manifests itself in a wider cross-judge distribution of first-round thresholds γ_{j1} (relative to the second round); it is likely that it also results in lower consistency.

With the above caveats, it is worthwhile to examine how inconsistency affects estimates of the marginal treatment effect arising from my data. The presence of inconsistency breaks the relationship between the approval rate of the judge and the identity of the judge's marginal claimants — there is instead a distribution of marginal claimants. Estimates of the marginal treatment effect under inconsistency thus average together individuals with different case strengths (**?**). I quantify the implications of inconsistency in two ways. First, I use my empirical setting to directly calculate how the estimated MTE of first-round approval on second-round approval changes with consistency. This approach is close to the data, and demonstrates that MTE estimates under inconsistency may be flatter than the underlying MTE. Second, I use the

method of **?** to calculate the estimated bias under different theoretical MTEs that might be encountered in other contexts.

A.2.6.1. MTE bias estimated from data. There is not a single way to quantify the effect of inconsistency on the estimated MTE. There are many potential non-degenerate joint distributions of first-round judge errors $\widetilde{\varepsilon}_{ij1}$that do *not* generate violations of monotonicity — for example, if all judges had the same error $\widetilde{\varepsilon}_{ij1}$ for each claimant. This is important because although the estimated parameters guarantee violations of monotonicity (recall from Section 2.4.3 that I bound the share of cases judge pairs disagree on above zero), different assumptions on the cross-judge joint distribution of $\widetilde{\varepsilon}_{ij1}$ allow for different counterfactuals that satisfy the monotonicity assumption. For simplicity I choose the most straightforward alternative: judges are perfectly consistent ($\sigma_{j1} = 0$ for all judges), guaranteeing montonicity is satisfied. I adjust thresholds γ_{j1} to keep the approval rate the same for all judges, then calculate the MTE under both the baseline coefficients and the counterfactual. Figure A9 shows that as consistency declines, the estimated MTE becomes shallower. This reflects how inconsistency generates a distribution of marginal claimants for each judge, each with a different treatment effect. Averaging over the distribution of marginal claimants for each judge reduces the cross-judge variation in the estimated MTE.

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

A.2.6.2. MTE bias for hypothetical MTEs. A second method to quantify the effect of inconsistency on estimated MTE comes from **?**. His approach is to take a second-order expansion of the MTE estimate around a baseline of perfect consistency, then study how estimates change. I

use the judge-specific estimates from the first round for the selection stage, and then estimate the bias under three different MTEs. I follow Klein and use the same functional form for the MTE as in his empirical example,

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

where I pick $\rho = \{0.5, 1, 1.5\}$. I assume that the true MTE is with respect to the refugee factor; $v = 1 - F_r(r_i)$. Panel A of Figure A10 plot these MTEs.

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

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

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

Panels B-D of Figure A10 plot estimates of the MTE bias at each point in the support of the instrument. As expected, the bias is worse in areas of the MTE with higher curvature, and worse for more-steeply sloped MTEs. The IV biases are 10, 19, and 26% for ρ of 0.5, 1, and 1.5. In this context, bias results more from the slope of the MTE (the second term in Equation A.14) than from curvature (first term).

Interestingly, and in contrast to the application-specific method in subsubsection A.2.6.1, the Klein method predicts negative bias over the support of the instrument. The cause of this difference is subtle. In subsubsection A.2.6.1, I kept overall approval rates the same while adjusting consistency σ_{j1}. Klein does not make this restriction, and so the results of Figure A10 partially reflect the addition of a large number of treated individuals. Because the distribution of underlying case strength r_i is right-tailed, the additional marginally treated claimants have disproportionately weak cases. This biases the estimated MTE downwards at all points.

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

A.2.7. Model parameters without additional regressors

The baseline model uses dummies for a late-week decision and whether the second-round hearing was made over lunch to aid in identification. In this section I present estimates from a model identified without regressors, as well as the main results. Identification now leans more strongly on functional form, though judge randomization still identifies relative consistency for judges with similar approval rates. Figure A12 presents the coefficients. The raw coefficients are similar to the baseline model but less precisely estimated. This reassuringly suggests that the results are driven mainly by the judge randomization (and to some degree the functional forms), but that the use of regressors additionally improves precision.

In Section 2.4.6 I present evidence that judicial inconsistency is due more to idiosyncratic observational errors than permanent ideological differences between judges in statutory interpretation. One fear with this approach is that errors arising from a late-week and noon-time decisions may be precisely idiosyncratic errors rather than ideological ones. In other words, the choice of regressors is determining the result. Table A17 tests the additional explanatory power of judge identity analogously to Table 4.9, but uses the no-regressor model probabilities. The results are comparable to the baseline specification: the model predicts second-stage approval well, but conditional on the model probabilities there is little additional predictive power from knowing the exact judge pairs. The distribution of the EB means of the judge pairs is very similar — in my preferred, rightmost specification, the standard deviation of the judge pair effects is 0.003 in the both models — and the F-test similarly does not reject that the judge pair effects are jointly zero.

In Table A18, I test the effect of experience and workload on inconsistency. Similarly to Table 4.10, I find that judges become dramatically more consistent after one year of experience,

but continue to make gains through at least the first ten years on the job. Higher caseloads decrease consistency (Columns 3 and 4), though only for judges with fewer than 6 years of experience (Column 5).

Table A19 contains estimates of the effect of judicial selection reform on judge consistency. As I describe in Section 2.3.3, changes to the laws governing judicial selections made it much more difficult for governments to grant judgeships to unqualified party supporters after 1988. In Section 2.4.9 I show that this reduced baseline estimates of consistency by approximately 75%. In the model estimated without regressors, the results are large but not quite so dramatic — in the baseline specification inconsistency declines by 0.7 from a pre-reform mean of 1.7.

Finally, Table A20 mirrors the results of the baseline model: estimates of judge thresholds γ_{js} are negatively correlated with survey measures of judge favorability to claimants, and model-estimated inconsistency σ_{j1} is negatively correlated with surveyed consistency. Again, this suggests that the correlation between model and survey results are not driven by the use of regressors in identification.

A.3. Appendix for *The Effects of Parental and Sibling Incarceration: Evidence from Ohio*

A.3.1. Using voter registry data to understand migration

Insofar as the goal of this paper is to provide the causal effects of intrafamily spillovers of incarceration, we are concerned about non-random migration out of study locations as a function of incarceration. If individuals with incarcerated parents are more likely to move out of state than those who parents were not incarcerated, then our positive results could be spurious: those children might be getting arrested and incarcerated at similar or even higher rates, but if it happened out of state, it would not be observed in our data. To understand whether migration out of our sample is a problem, we employ voter registry data to answer two questions. First, does the study sample migrate to other areas of Ohio outside of our study sample? In the case of crime and education outcomes, we are limited to viewing the three largest counties in Ohio: Cuyahoga, Franklin, and Hamilton. That concern is absent for teen pregnancy or fertility effects as our data spans the entire state. Second, does the study population migrate out of the state?

Voter registration data is a useful way to test both of these concerns since the registry contains adult address information, so we can track children with incarcerated parents as adults. We observe the address of anyone who was ever registered to vote in Ohio between June 2000 and November 2016, approximately 11.4 million unique individuals. The inclusion of an individual in the registry provides evidence that the person is living in Ohio. Comparing an individual's birth address to their voter registry address shows us whether the individual has moved to other parts of the state. Unlike other states that preclude some ex-convicts from voting, Ohio only restricts convicted felons from voting or being part of the voter registry during their time in prison (Ohio state code 2961.01). Anyone granted parole, judicial release, or a conditional pardon or

is released under a non-jail community control sanction or a post-release control sanction is eligible to be a voter, and hence can register as a voter. Therefore, in Ohio, the voter registry is one of the best places to find current address, and hence, will be our measure of adult address. We link voter registry data to our birth certificate data using name and date of birth. We construct distance between birth addresses and voter registration address in kilometers using flight path distance.

Looking first at whether incarceration induces migration out of state, **??** provides evidence to the contrary. Incarceration leads to an insignificant positive effect on being registered as a voter in Ohio in the full sample. The effects for boys is larger than girls by about 2.5 percentage points, but both are insignificant and positive. Overall, we see a large share of our study population, 69%, are registered as voters, and hence clearly observable as living within the state. A large share of the non-registered are likely to be residing in Ohio, but are simply not registered voters. In future iterations of this paper, we will match our study sample to the neighboring states to provide further evidence on migration.

Looking to columns (4-6), we see that the incarceration of a parent leads to a significant increase of 11 percentage points in voter registration for the quartile of our study sample born in the poorest neighborhoods. The majority of this effect is coming from boys with incarcerated parents. This is the exact opposite of the effect we observe for their criminal behavior, where the crime reductions are largest for boys coming from the lowest socioeconomic status neighborhoods. Taken together, this is compelling evidence that the reductions in crime associated with parental incarceration are not driven by differential out-migration from the state. Instead, if children with incarcerated are more likely to remain in state, this suggests that our estimates

may even be downward biased. It may also be that having an incarcerated parent acts to mobilize the young to vote for new policies, but the opposite directions of the crime and voting registry effects make it implausible that out-of-state migration could be driving our positive results.

?? and **??** show that parental incarceration also does not lead to out-migration within state. Restricting the analysis sample to those who were observed in the voter registry data, **??** shows that parental incarceration leads to no statistically or economically significant effects on the distance between a child's birth address and voter registry address. For boys with female parents, who had the largest negative impacts on their likelihood of being in court or incarcerated because of parental incarceration, we find an insignificant reduction in lifetime migration distance of about 10% relative to the dependent variable mean.

?? uses county of residence in the voting data to see whether children of incarcerated parents are more likely to move away from Cuyahoga, Franklin, and Hamilton counties. Among those children in our data who had incarcerated parents and are registered to vote in Ohio, around 75% live in one of these three counties. Migration rates for this population are generally low, with around 80% living in their county of birth.[8] Having an incarcerated parent does not affect either the probability that a child moves from their birth county or that they are no longer a resident of the counties for which we observe outcomes. Therefore, out-migration from the geographic study window is unlikely to generate the negative impacts on the criminal outcomes measured here.

The proceeding results demonstrate that out-migration is unlikely to be driving the reductions in contact with the criminal justice system seen in our main results. However, the increase

[8]If anything, this is probably an overestimate of migration rates for children in our sample since those individuals who register to vote are likely to be more affluent and geographically mobile.

in voter registration for those coming from the poorest neighborhoods is interesting in its own right and could result from three forces. First, it could represent a reduction in migration for children with incarcerated parents as a result of various sentencing restrictions like parole requirements to stay within city limits given by the judge or due to other decisions within the family. Second, it could result from children being incarcerated during the sample period and thus lacking a voting address. Since children of non-incarcerated parents are more likely to be incarcerated, this could explain some of the effect. Third, it could mark an increase in activism on the part of children with incarcerated parents. We don't take a stand on any of these forces, and as is usually true, the answer is likely that they are all at play. Incorporating data from neighboring states will be helpful in further establishing that differential outmigration is not biasing the results.

A.3.2. Incarceration effects on family size

To further characterize the impacts of incarceration on defendants and to understand how incarceration potentially affects within family resource allocation, we analyze the impacts of incarceration on defendants' future fertility outcomes. For defendants, incarceration could affect the likelihood of having children through "incapacitation effects," as they are much less likely to have children while physically incarcerated, and through their ability or desire to have children after release. In particular, formerly incarcerated individuals may find it more difficult to find a mate with whom to have children or may separate from pre-existing partners due to stigma and other negative impacts of incarceration, such as employment loss. Even when incarceration does not make it more difficult to find a partner, earnings losses may make it harder to raise a children, thereby reducing the likelihood for an individual to choose to have more children. For their children at the time of incarceration, changes to family fertility potentially affect the social or economic support received by the other family members. These impacts would likely be relatively small. Although there is a large literature in economics on trade-offs between child quantity and quality, increasing family size does not seem to reduce important observable outcomes such as child educational attainment in developed countries (**?**).

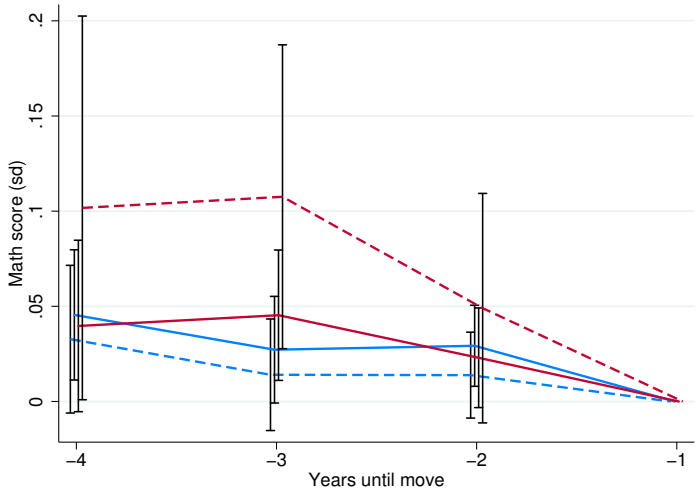
?? analyzes the cumulative number of children born to a defendant in each quarter after being charged. Incarceration leads to sustained, albeit small, negative effects on family size. The figure shows coefficients from a quarter-by-quarter regression of the cumulative number of new children since judge assignment on the leave-out judge severity. Ten quarters after incarceration, the incarcerated have an average of 0.015 fewer children than those who were not incarcerated. This number grows to .023 fewer children after 20 quarters. With reduced family sizes, it is possible for incarcerated parents to reallocate their parenting resources towards

their pre-existing children away from the counterfactual children they would have had. This is unlikely to be a major driver of the results we see. The typical defendant has on average 0.61 children, hence these impacts are miniscule in proportion. The results still do have implications for the incarcerated population as a whole, who may have preferred to have more children. The lack of either ability or desire could either be co-moving or driving some of the increased criminogenic impacts observed in the literature and in Figure 4.25.

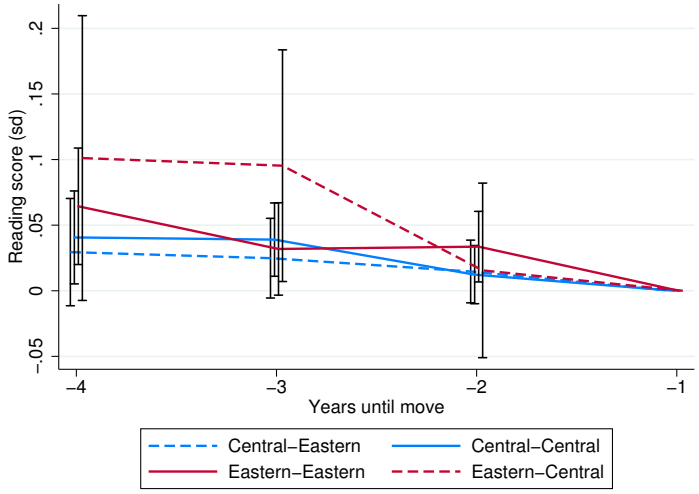
A.4. Appendix Tables and Figures

Figure A1. Pre-move trends in academic outcomes, by mover type without additional controls

(a) Math scores

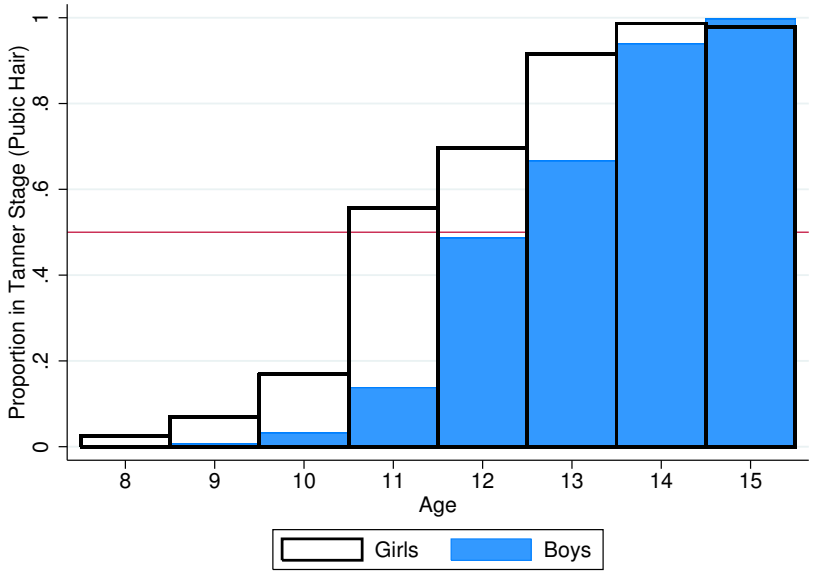


(b) Reading scores



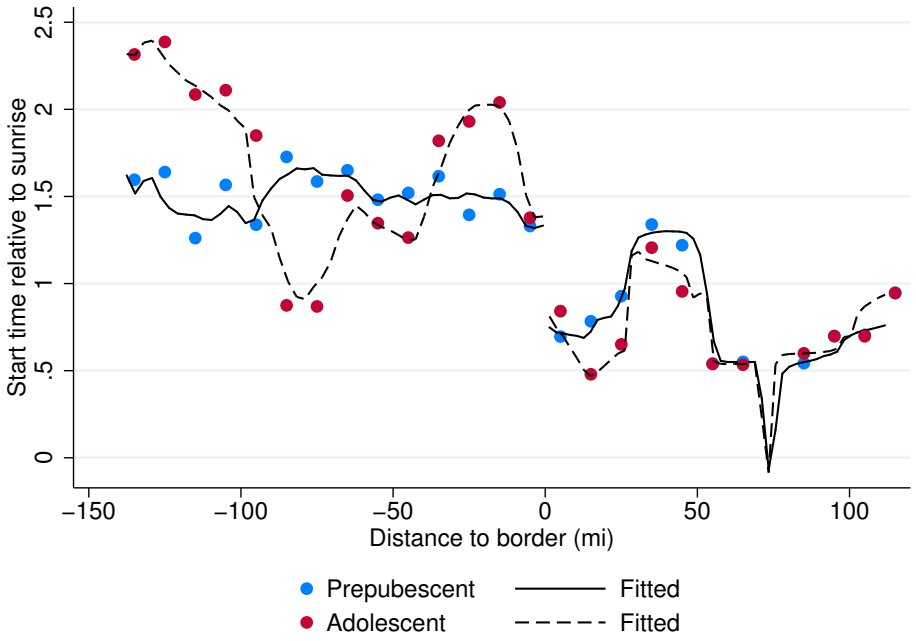
Displays the pre-move achievement trends for the four years leading up to a move of 25 miles. Results reported separately for four groups of movers: within CT, within ET, ET to CT, and CT to ET. Coefficients recovered from a regression of test scores on time-until-move dummies and a fixed effect for the period before the move. Standard errors are clustered at the individual level, and included as bars representing 95% confidence intervals.

Figure A3. Tanner stage 3 proportions by age and sex



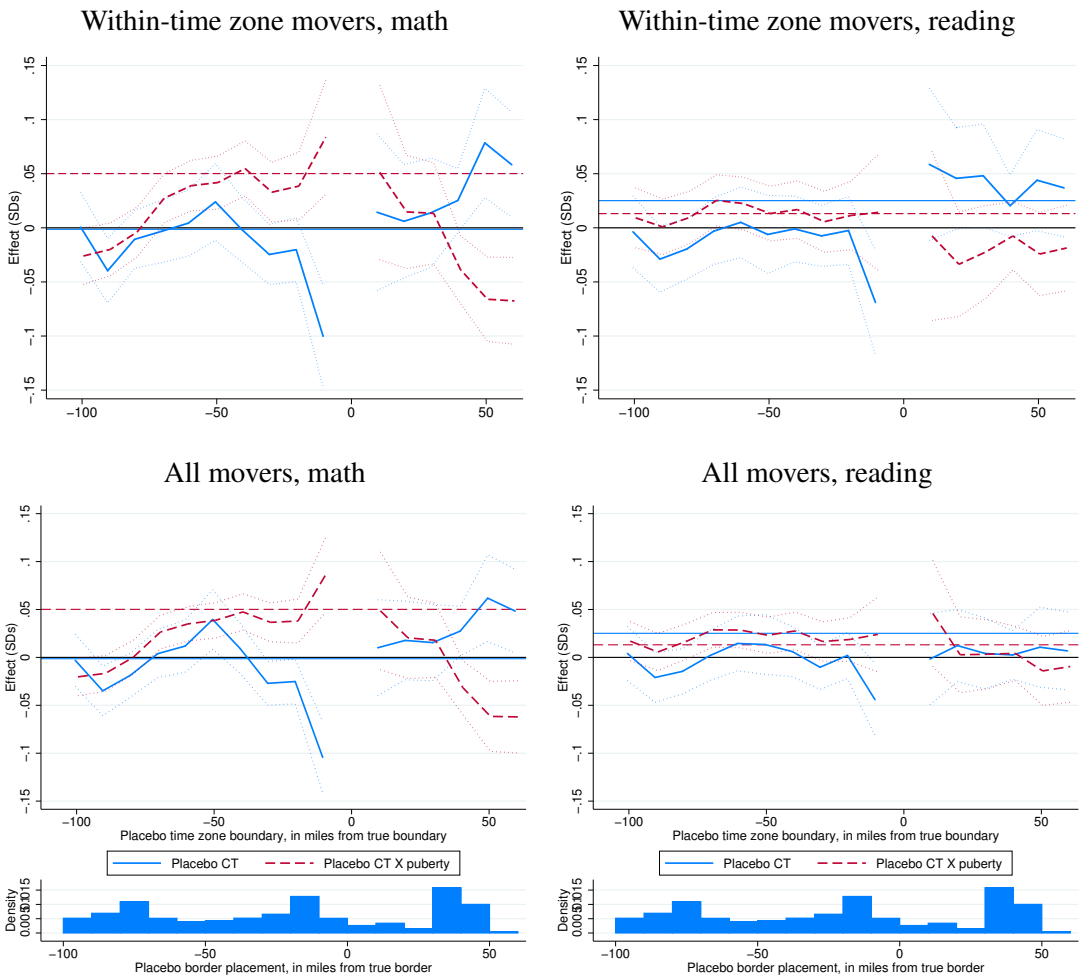
Displays proportion of children who had entered the Tanner Stage for pubic hair development at a given age for males and females. Horizontal line represents median child entering the stage.

Figure A4. Relative start time near the time zone boundary



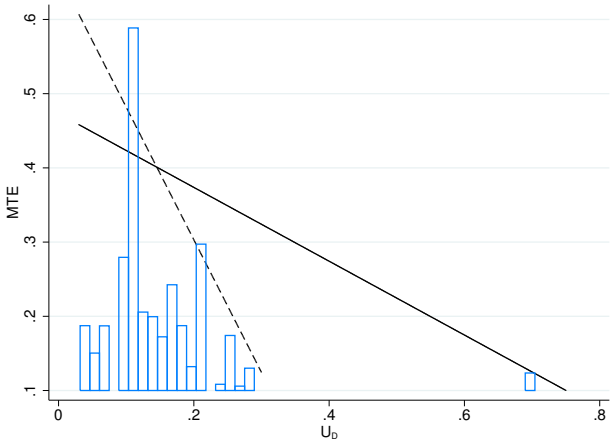
Displays a nonparametric regression of relative start time (start time minus sunrise) on distance to the time zone boundary, estimated separately for each time zone. Scatter points are ten mile bin averages.

Figure A5. Effect of placebo time zones on academic achievement, no sample exclusion near true time zone boundary



Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Thin horizontal lines represent baseline coefficient estimates. We generate placebo time zones in ten mile increments from the true time zone boundary. Then, placebo coefficients are calculated from individual regressions of the outcome on the true time zone interacted with puberty, and the placebo time zone interacted with puberty. All specifications include age-gender dummies, longitude controls, school demographic means (FRL, male, black, Asian, and Hispanic) and individual fixed effects. Standard errors clustered at the individual level. We display results including and excluding cross-time zone movers.

Figure A7. MTE of second-round approval on first-round approval judge



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

Figure A8. Scatter plot of first- and second-round consistency σ_{js}

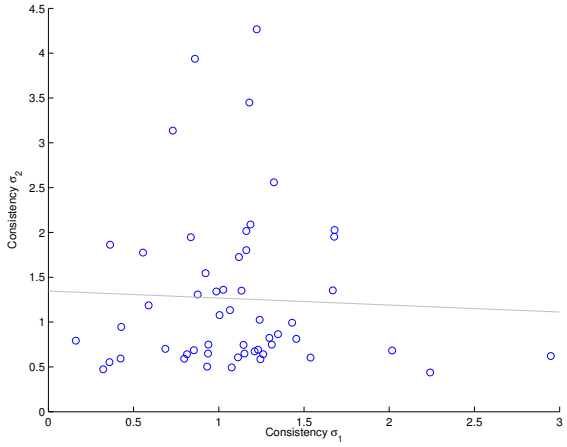


Figure A9. Estimated MTE at baseline and under consistency

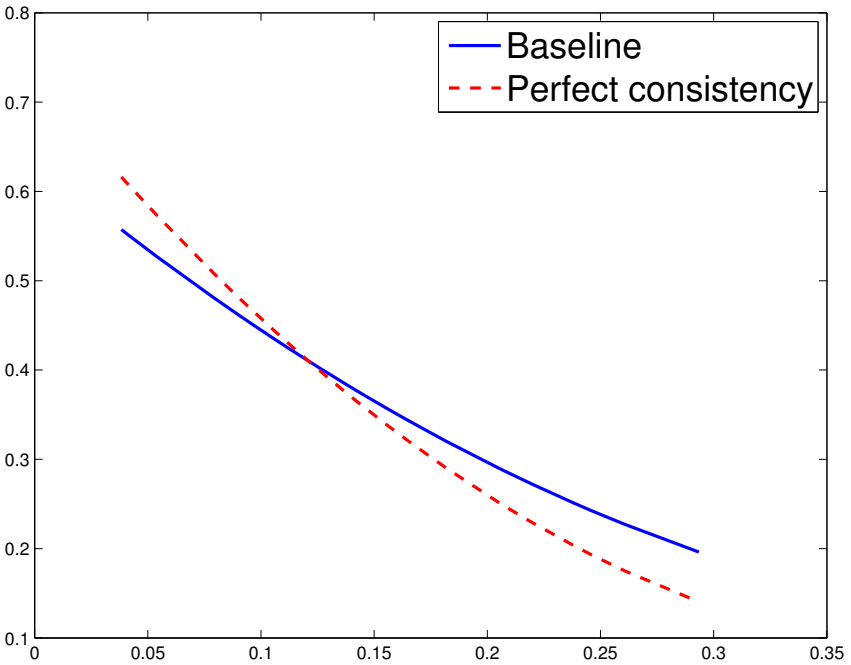
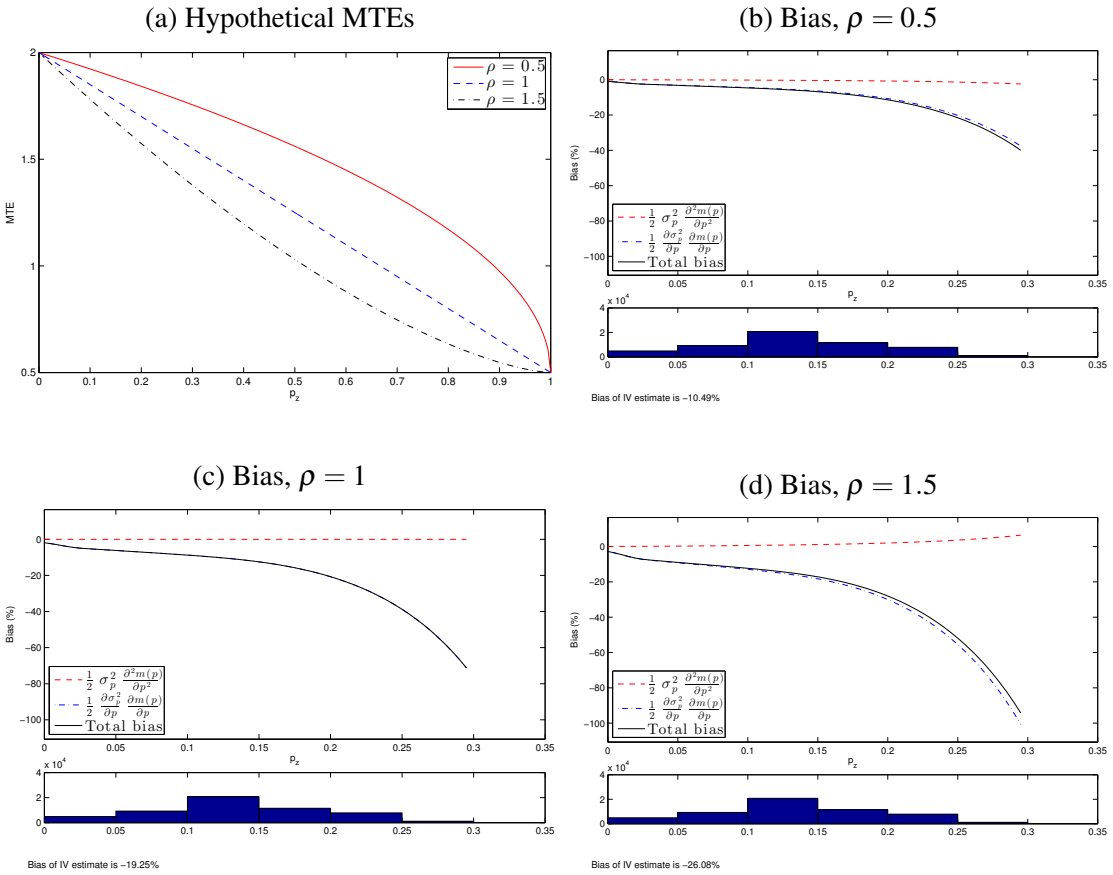


Figure demonstrates how inconsistency affects estimate of MTE. I plot the baseline MTE in solid blue, then use model estimates to construct an estimate of the MTE if all judges were perfectly consistent (ie, $\sigma_{j1} = 0$) but had the same average approval rate.

Figure A10. Bias for different MTEs

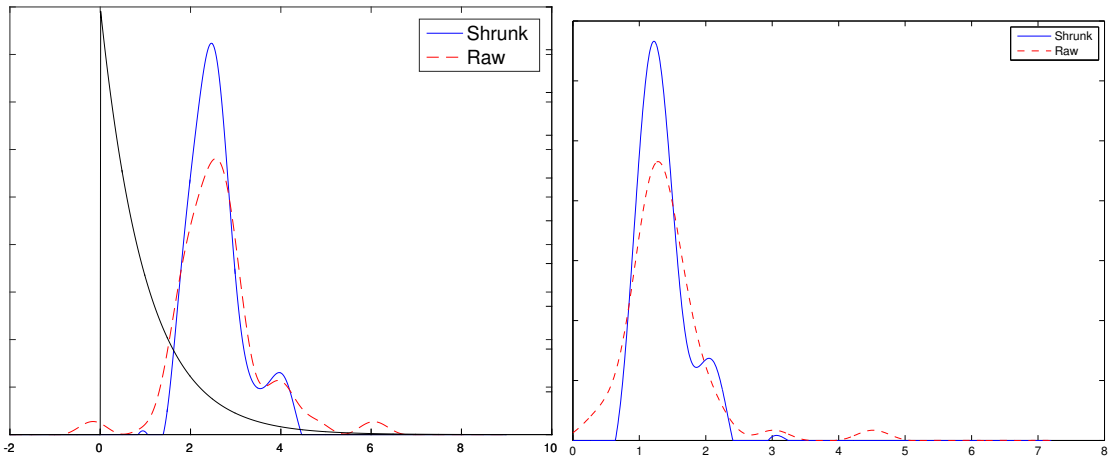


Panel A plots the marginal treatment effects $m(v) = 5 + 1.5 * (1-v)^{\rho}$ for $\rho = \{0.5, 1, 1.5\}$. Panels B, C and D plot bias over the support of the main mass of the instrument for each MTE.

Figure A12. Distribution of judge coefficients, model identified without regressors

(a) Threshold γ_1, first round

(b) Observational error σ_1, first round



(c) Threshold γ_2, second round

(d) Observational error σ_2, second round

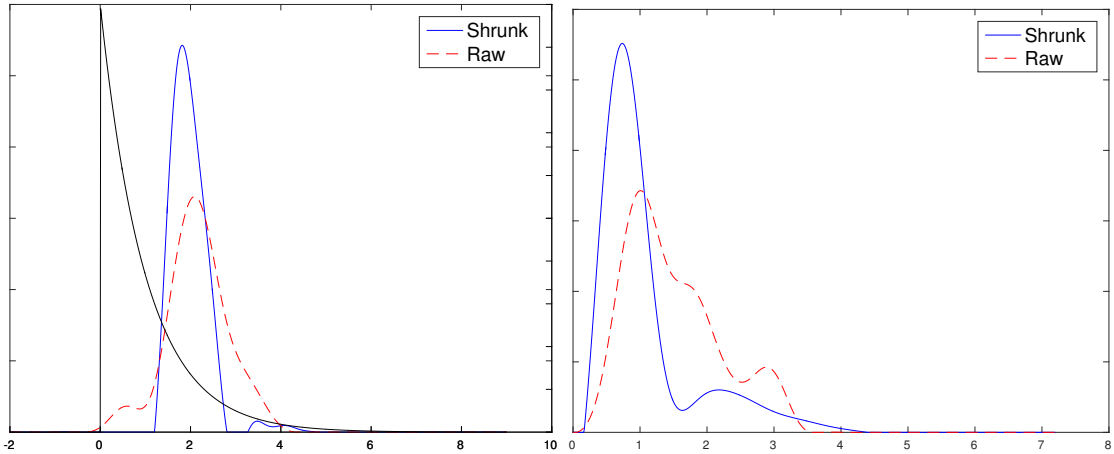


Figure displays coefficients for decision model $\mathbb{1}[r_i > \gamma_{js} + \widetilde{\varepsilon}_{ijs}]$, $\widetilde{\varepsilon}_{ijs} \sim \mathcal{N}(0, \sigma^2_{js})$. All models allow the parameters of the Pareto distribution of r_i to vary flexibly by office of origination as well as after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

Table A1. Academic outcomes on school start time for varying mover definitions, with student fixed effects

	Math (SDs)					Reading (SDs)				
	(1) dist	(2) 15 mi	(3) 20 mi	(4) 25 mi	(5) 30 mi	(6) dist	(7) 15 mi	(8) 20 mi	(9) 25 mi	(10) 30 mi
Start time - sunrise (h)	0.037	0.029	0.014	0.009	0.009	0.037	0.034	0.026	0.061*	0.053
	(0.034)	(0.033)	(0.031)	(0.035)	(0.037)	(0.036)	(0.034)	(0.032)	(0.036)	(0.038)
Start time X puberty	0.036**	0.038**	0.070***	0.073***	0.060***	0.007	0.011	0.018	-0.004	-0.008
	(0.018)	(0.017)	(0.018)	(0.019)	(0.022)	(0.019)	(0.018)	(0.018)	(0.020)	(0.023)
P(Start+Start X puberty=0)	0.001	0.002	0.000	0.001	0.004	0.029	0.025	0.033	0.014	0.049
Cragg-Donald F-stat	610.14	611.40	677.49	542.01	542.98	684.27	701.42	766.47	619.26	612.31
Number of students	33712	35744	28969	24768	21557	34144	36197	29393	25191	21957
Observations	143921	153462	120233	99835	84165	150800	160997	126110	104791	88408

Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Start time and its interaction with puberty are instrumented by time zone and the interaction of time zone and puberty. All specifications include age-gender dummies, longitude controls, school demographic means (FRL, male, black, Asian, and Hispanic), and individual fixed effects. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A2. Academic and behavioral outcomes on start time, with student fixed effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Panel A: Math Test Scores (SDs)															
Start time - sunrise (h)	0.012	0.011	0.009	0.020	0.028	0.028	0.031	0.036	0.037	0.009	0.014	0.014	0.003	0.012	0.012
	(0.041)	(0.035)	(0.035)	(0.036)	(0.037)	(0.036)	(0.035)	(0.036)	(0.036)	(0.035)	(0.034)	(0.034)	(0.036)	(0.034)	(0.034)
Start time X puberty	0.073***	0.051***	0.054***	0.057***	0.037*	0.039*	0.065***	0.041**	0.043**	0.073***	0.050***	0.053***	0.076***	0.050***	0.053***
	(0.020)	(0.019)	(0.019)	(0.021)	(0.021)	(0.021)	(0.020)	(0.020)	(0.020)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
P(Start+Start X puberty=0)	0.002	0.005	0.005	0.001	0.003	0.002	0.000	0.001	0.000	0.001	0.003	0.002	0.001	0.004	0.003
Cragg-Donald F-stat	405.14	593.76	604.37	588.76	593.74	606.81	580.48	599.33	612.93	542.01	640.62	655.01	534.48	638.53	655.08
Number of students	24768	23516	23516	24768	23516	23516	24545	23294	23294	24768	23516	23516	24765	23514	23514
Observations	99835	91853	91853	99835	91853	91853	98751	90852	90852	99835	91853	91853	99823	91846	91846
Panel B: Reading Test Scores (SDs)															
Start time - sunrise (h)	0.087**	0.061*	0.061*	0.081**	0.075**	0.074**	0.071**	0.065*	0.065*	0.061*	0.049	0.048	0.051	0.046	0.046
	(0.041)	(0.035)	(0.035)	(0.037)	(0.037)	(0.036)	(0.035)	(0.036)	(0.035)	(0.036)	(0.034)	(0.034)	(0.036)	(0.034)	(0.034)
Start time X puberty	-0.013	-0.009	-0.008	-0.023	-0.022	-0.022	-0.011	-0.013	-0.013	-0.004	-0.003	-0.003	0.000	-0.002	-0.002
	(0.021)	(0.020)	(0.020)	(0.022)	(0.021)	(0.021)	(0.021)	(0.020)	(0.020)	(0.020)	(0.019)	(0.019)	(0.020)	(0.019)	(0.019)
P(Start+Start X puberty=0)	0.004	0.015	0.015	0.008	0.014	0.014	0.008	0.015	0.014	0.014	0.027	0.027	0.025	0.030	0.030
Cragg-Donald F-stat	486.65	679.86	687.26	637.22	648.26	671.04	656.76	675.89	697.05	619.26	729.44	746.01	616.60	725.65	742.75
Number of students	25191	24048	24048	25191	24048	24048	24963	23823	23823	25191	24048	24048	25189	24045	24045
Observations	104791	96788	96788	104791	96788	96788	103547	95641	95641	104791	96788	96788	104776	96776	96776
Panel C: Absence Rates															
Start time - sunrise (h)	-1.860***	-1.463***	-1.431***	-0.718	-0.709	-0.695	-0.848*	-0.789*	-0.772*	-0.869*	-0.874*	-0.859*	-0.965**	-0.904*	-0.880*
	(0.590)	(0.505)	(0.502)	(0.474)	(0.483)	(0.479)	(0.460)	(0.471)	(0.467)	(0.485)	(0.467)	(0.464)	(0.492)	(0.470)	(0.466)
Start time X puberty	0.857***	0.677**	0.637**	0.395	0.330	0.304	0.439	0.353	0.320	0.469*	0.384	0.365	0.491*	0.396	0.367
	(0.294)	(0.278)	(0.275)	(0.285)	(0.286)	(0.283)	(0.274)	(0.278)	(0.275)	(0.268)	(0.268)	(0.265)	(0.269)	(0.270)	(0.266)
P(Start+Start X puberty=0)	0.010	0.012	0.011	0.274	0.182	0.166	0.156	0.117	0.103	0.219	0.091	0.087	0.151	0.081	0.077
Cragg-Donald F-stat	274.18	413.70	416.25	425.38	431.70	439.86	453.02	458.47	467.24	383.62	451.74	458.44	373.38	447.12	454.86
Number of students	15906	15130	15130	15906	15130	15130	15906	15130	15130	15906	15130	15130	15903	15128	15128
Observations	66263	61128	61128	66263	61128	61128	66263	61128	61128	66263	61128	61128	66252	61122	61122
Urban and log income	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Size and S/T ratio	No	No	Yes	No	No	Yes	No	No	Yes	No	No	Yes	No	No	Yes
District controls	No	No	No	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No
District grade 3 controls	No	No	No	No	No	No	Yes	Yes	Yes	No	No	No	No	No	No
School controls	No	No	No	No	No	No	No	No	No	Yes	Yes	Yes	No	No	No
School-grade controls	No	No	No	No	No	No	No	No	No	No	No	No	Yes	Yes	Yes

Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Absentee rate is the fraction of days the child missed school. Start time and its interaction with puberty are instrumented by time zone. Sample is all children who moved more than 25 miles. All specifications include age-gender dummies, longitude, and individual fixed effects. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3. Outcomes on school start time, with latitude and school test grade scores

	Math			Reading		
	(1)	(2)	(3)	(4)	(5)	(6)
Start time - sunrise (h)	0.009	-0.035	0.015	0.061*	0.035	0.051
	(0.035)	(0.033)	(0.037)	(0.036)	(0.034)	(0.037)
Start time X puberty	0.073***	0.085***	0.073***	-0.004	0.004	-0.001
	(0.019)	(0.019)	(0.020)	(0.020)	(0.020)	(0.020)
Latitude controls	No	Yes	No	No	Yes	No
Third grade district scores	No	No	Yes	No	No	Yes
P(Start+Start X puberty=0)	0.001	0.029	0.001	0.014	0.069	0.035
Cragg-Donald F-stat	542.01	631.95	508.46	619.26	715.55	589.27
Number of students	24768	24768	24288	25191	25191	24730
Observations	99835	99835	97483	104791	104791	102276

Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Start time and its interaction with puberty are instrumented by time zone and the interaction of time zone and puberty. Sample is all children who moved more than 25 miles. All specifications include age-gender dummies, longitude controls, school demographic means (FRL, male, black, Asian, and Hispanic), and individual fixed effects. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4. Florida school and peer characteristics on move

	% FRL	% male	% black	% Hispanic	% Asian	S/T	Med income
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Move, ET-ET	-4.494***	-0.452***	0.186	-0.100	0.263***	0.258***	1010.277*
	(0.726)	(0.118)	(0.801)	(0.224)	(0.059)	(0.081)	(601.359)
Move, CT-CT	-1.681***	-0.316***	-0.582**	0.110***	-0.011	0.190***	-429.606***
	(0.280)	(0.054)	(0.227)	(0.037)	(0.025)	(0.038)	(162.849)
Move, ET-CT	0.115	-0.009	-15.350***	0.025	0.426***	0.124	-4778.338***
	(0.923)	(0.162)	(1.015)	(0.183)	(0.084)	(0.103)	(731.901)
Move, CT-ET	-4.513***	-0.557***	13.965***	0.495***	0.023	0.113	5729.001***
	(0.939)	(0.163)	(1.010)	(0.166)	(0.088)	(0.101)	(752.117)
P(ET-CT=CT-ET)	0.002	0.029	0.000	0.105	0.003	0.944	0.000
Observations	31763	31763	31763	31763	31763	31763	27747

Dependent variable as noted in panel heading. Regression is of school/zip summary stat on move, with student X moving event FE. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5. Alternative definitions of puberty

	Math (SDs)				Reading (SDs)			
	(1) Preferred	(2) Stage 2	(3) Stage 4	(4) BG	(5) Preferred	(6) Stage 2	(7) Stage 4	(8) BG
Start time - sunrise (h)	0.009	0.011	0.032	0.025	0.061*	0.057	0.056	0.058
	(0.035)	(0.036)	(0.035)	(0.035)	(0.036)	(0.036)	(0.036)	(0.036)
Start time X puberty	0.073***	0.064***	0.029	0.040**	-0.004	0.003	0.006	0.002
	(0.019)	(0.019)	(0.020)	(0.019)	(0.020)	(0.020)	(0.021)	(0.020)
P(Start+Start X puberty=0)	0.001	0.003	0.005	0.008	0.014	0.012	0.002	0.010
Cragg-Donald F-stat	542.01	566.32	444.15	542.35	619.26	655.35	487.58	615.52
Number of students	24768	24768	24768	24768	25191	25191	25191	25191
Observations	99835	99835	99835	99835	104791	104791	104791	104791

Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Absentee rate is the fraction of days the child missed school. Start time and its interaction with puberty are instrumented by time zone and the interaction of time zone and puberty. Sample is all children who moved more than 25 miles. All specifications include age-gender dummies, longitude controls, school demographic means (FRL, male, black, Asian, and Hispanic) and individual fixed effects. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6. Academic and behavioral outcomes on start time, with student fixed effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: First stage, relative start time (hours)							
CT (=1)	0.598***	0.475***	0.585***	0.547***	0.475***	0.584***	0.547***
	(0.015)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
Observations	115778	115778	115778	115778	115778	115778	115778
Panel B: IV estimates, math test scores (SDs) on relative start time							
Start time - sunrise (h)	-0.005	0.047	0.048*	0.044	0.045	0.048*	0.043
	(0.019)	(0.034)	(0.028)	(0.029)	(0.034)	(0.028)	(0.029)
Cragg-Donald F-stat	2254.173	744.796	1120.532	1002.330	746.364	1120.434	1003.020
Panel C: IV estimates, reading test scores (SDs) on relative start times							
Start time - sunrise (h)	0.061***	0.081**	0.069**	0.059**	0.080**	0.069**	0.059**
	(0.019)	(0.032)	(0.028)	(0.028)	(0.032)	(0.028)	(0.028)
Cragg-Donald F-stat	2587.05	911.72	1209.23	1151.57	913.31	1209.80	1152.03
Panel D: IV estimates, absence rate (%) on relative start times							
Start time - sunrise (h)	-0.664**	-1.539***	-0.549	-0.670*	-1.510***	-0.559	-0.672*
	(0.275)	(0.501)	(0.391)	(0.407)	(0.499)	(0.389)	(0.405)
Longitude	No	Yes	Yes	Yes	Yes	Yes	Yes
District quality	No	No	Yes	No	No	Yes	No
School quality	No	No	No	Yes	No	No	Yes
Time since move	No	No	No	No	Yes	Yes	Yes
Cragg-Donald F-stat	1394.52	475.67	721.91	669.77	476.44	722.82	669.98

Dependent variable as noted in panel heading. Test scores measured in SDs normalized at the grade-year level for the entire state. Absentee rate is the fraction of days the child missed school. Relative start time instrumented by time zone. Sample is all children who moved more than 25 miles. All specifications include age-gender dummies and individual fixed effects. Sample size is fixed within panels: 34018 students and 115778 student-years in Panel A, 24768 students and 99835 student-years in Panel b, 25191 students and 104791 student-years in Panel C, and 15906 students and 66263 student-years in Panel D. Standard errors in parentheses and clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A7. Hours of sleep by time zone

	(1)	(2)	(3)
Central	0.081	0.103	
	(0.088)	(0.131)	
Puberty	-0.451***	-0.804***	-0.676***
	(0.055)	(0.122)	(0.134)
Weekend	1.421***	1.192***	1.229***
	(0.102)	(0.158)	(0.158)
Central X weekend	-0.107	-0.166	-0.102
	(0.156)	(0.194)	(0.188)
Central X puberty	0.218	0.183	0.257
	(0.139)	(0.185)	(0.195)
Weekend X puberty	0.384***	0.616***	0.586***
	(0.087)	(0.161)	(0.150)
Central X wkend X puberty	-0.215	-0.149	-0.229
	(0.168)	(0.239)	(0.224)
P(Central + Central X weekend = 0)	0.830	0.566	
P(Central + Central X puberty = 0)	0.074	0.085	
Demographic controls	No	Yes	No
Student fixed effects	No	No	Yes
Observations	6,084	3,737	6,084

Dependent variable is hours of sleep per night. Sample is all children 6-19 in the Child Development Supplement of the Panel Study of Income Dynamics within 400 miles of the ET-CT time zone boundary in a state with a single time zone. Demographic controls in Column 2 include gender, race, and FRL status. Standard errors in parentheses and clustered at the state level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A8. Judge summary statistics

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

Table A9. Randomization using name-imputed continent of origin

	Male	Africa	Asia	South America	Predicted approval	1st-round mean approval
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: First round judges						
First-round approval rate	-0.002	-0.068***	-0.011	0.098	-0.002	
	(0.020)	(0.023)	(0.040)	(0.071)	(0.002)	
F-stat	0.87	2.90	3.04	6.65	3.51	
Prob	0.75	0.00	0.00	0.00	0.00	
Observations	58,604	58,604	58,604	58,604	58,604	
Panel B: Second round judges						
Second-round approval rate	-0.042	0.032	-0.022	0.027	0.002	-0.027
	(0.033)	(0.039)	(0.042)	(0.041)	(0.003)	(0.018)
F-stat	1.07	1.83	1.19	1.61	1.54	4.02
Prob	0.33	0.00	0.15	0.00	0.01	0.00
Observations	8,446	8,446	8,446	8,446	8,446	8,446

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

Table A10. Testing effect of regressors on distribution of judge errors

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

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

Table A11. Lawyer characteristics, survey respondents vs lawyer population

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

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

Table A12. Model coefficients on survey responses including accuracy response

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

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

Table A13. Inconsistency σ_1 on survey responses with approval controls

	(1)	(2)	(3)	(4)	(5)	(6)
Favorability, SD	0.0408	-0.108			0.125**	-0.0430
	(0.0694)	(0.105)			(0.0539)	(0.0981)
Inconsistency, SD			0.166***	0.151**	0.210***	0.131***
			(0.0563)	(0.0583)	(0.0505)	(0.0437)
Approval rate control	Yes	Yes	Yes	Yes	Yes	Yes
Observations	136	136	136	136	136	136

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

Table A14. Approval rate for judges before and after reform

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

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

Table A15. Inconsistency before and after reform, control for approval rate

	Baseline			Experience control in σ_1		
	(1)	(2)	(3)	(4)	(5)	(6)
After 1988 reform (=1)	-0.211	-0.362	-0.379	-1.706***	-1.772***	-1.717**
	(0.139)	(0.251)	(0.260)	(0.566)	(0.627)	(0.647)
Liberal appointee (=1)			-0.0429			-0.143
			(0.104)			(0.119)
Male judge (=1)			-0.201*			-0.216*
			(0.113)			(0.128)
Year appointed	No	Yes	Yes	No	Yes	Yes
Pre-reform mean	1.26	1.26	1.26	2.20	2.20	2.20
N judges	53	53	53	53	53	53

Estimated with Hanushek (1974) correction for estimated dependent variable and control for judge approval rate. Dependent variable is consistency σ_{j1}, which is estimated from decision model $\mathbb{1}[r_i > X_{ijs}\beta_s + \gamma_{js} + \widetilde{\varepsilon}_{ijs}(W_{ijs})]$, $\widetilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\widetilde{\sigma}_{js} + W_{ijs}\psi_s)$. In the right-hand panel, β_s and ψ_s include dummies for more than 1, 5, and 10 years of experience. Robust standard errors in parentheses and clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A16. Second-round inconsistency for judges before and after reform

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

Estimated with Hanushek (1974) correction for estimated dependent variable. Dependent variable is consistency σ_{j2}, which is estimated from decision model $\mathbb{1}[r_i > X_{ijs}\beta_s + \gamma_{js} + \widetilde{\varepsilon}_{ijs}(W_{ijs})]$, $\widetilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\widetilde{\sigma}_{js} + W_{ijs}\psi_s)$. In the right-hand panel, β_s and ψ_s include dummies for more than 1, 5, and 10 years of experience. Robust standard errors in parentheses and clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A17. Second-round outcome on model approval probability and judge-pair FEs

	Judge-pair round FEs		Judge-pair FEs	
	(1)	(2)	(3)	(4)
Model approval probability	0.949***	0.947***	0.976***	0.975***
	(0.166)	(0.166)	(0.0478)	(0.0478)
Model controls	No	Yes	No	Yes
Mean approval	0.44	0.44	0.44	0.44
F-stat for judge pairs	1.02	1.02	0.98	0.99
P-value	0.318	0.314	0.641	0.610
Bootstrap p-value	0.674	0.657	0.834	0.804
SD of judge-pair EB means	0.006	0.006	0.004	0.003
Observations	8,196	8,196	8,196	8,196

Regresses second-round approval on model-predicted likelihood of approval and judge-pair fixed effects. In contrast to Table 4.9, model is estimated without using regressors for identification. Left two columns construct judge-pair FEs accounting for order of assignment; right two columns ignore this distinction. Model controls include office of origination, pre-post 2002, and an end-of-week and noon hearing dummy for the second-round hearing. Standard errors clustered at the judge level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A18. First-round judge consistency by experience and workload

	(1)	(2)	(3)	(4)	(5)
Coefficients ψ affecting judge inconsistency σ_1					
Experience $>$ 1 year	-0.797^{**}	-0.600^{***}		-0.557^{**}	-0.462^{**}
	(0.365)	(0.212)		(0.267)	(0.203)
Experience $>$ 5 years	-0.351^{***}	-0.345^{*}		-0.049	-0.054
	(0.104)	(0.199)		(0.561)	(0.380)
Experience $>$ 10 years	-0.530	-0.452^{***}		-0.447^{***}	-0.501^{***}
	(0.524)	(0.101)		(0.170)	(0.012)
Log caseload			0.204^{***}	0.139^{***}	
			(0.020)	(0.038)	
Log caseload ($\leq$ 5 yrs exp)					0.179^{***}
					(0.025)
Log caseload ($>$ 5 yrs exp)					0.056
					(0.146)
SD of γ_1	0.817	0.715	1.417	1.122	1.044
SD of σ_1	1.562	1.486	0.537	0.991	1.084
Second-round experience control	Yes	Yes	No	Yes	Yes
Career number of cases	No	Yes	No	No	No

Reports coefficients for decision model $\mathbb{1}[r_i > \gamma_{js} + X_{ijs}\beta_s + \widetilde{\varepsilon}_{ijs}(W_{ijs})]$, $\widetilde{\varepsilon}_{ijs} \sim \mathcal{N}(0, \sigma^2_{js})$, $\sigma_{js}(W_{ijs}) = e^{\widetilde{\sigma}_{js}}$. In contrast to the baseline model, the reported models do not use timing regressors for identification. All models include controls for time/date of decision in β, and allow the parameters of the Pareto distribution of r_i to vary flexibly by office of origination as well as after 2002 relative to before 2002, when there were legislative changes that may have impacted the distribution of case quality. Standard errors clustered at the level of the first stage judge. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A19. Inconsistency before and after judge selection reform

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

Estimated with Hanushek (1974) correction for estimated dependent variable. Dependent variable is consistency σ_{j1}, which is estimated from decision model $\mathbb{1}[r_i > \gamma_{js} + X_{ijs}\beta_s + \widetilde{\varepsilon}_{ijs}(W_{ijs})]$, $\widetilde{\varepsilon}_{ijs}(W_{ijs}) \sim \mathcal{N}(0, \sigma_{js}(W_{ijs})^2)$, $\sigma_{js}(W_{ijs}) = \exp(\widetilde{\sigma}_{js} + W_{ijs}\psi_s)$. In contrast to the baseline model, the reported models do not use timing regressors for identification. In the right-hand panel, β_s and ψ_s include dummies for more than 1, 5, and 10 years of experience. Robust standard errors in parentheses and clustered at the judge level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A20. Model coefficients on survey responses

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

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