The Short- and Long-Run Effects of Gender Discrimination in the Courthouse*

Ozkan Eren

University of California, Riverside

Randi Hjalmarsson

University of Gothenburg

Orgul Ozturk

University of South Carolina

March 2025

Abstract

We take advantage of conditionally random judge assignment in a Southern U.S. state to study the effects of same-gender judge assignment on both high-stakes juvenile justice decisions and short- and long-run socioeconomic outcomes. Using a generalized differences-in-differences approach, we find evidence of in-group gender gaps in sentencing decisions: relative to males, female judges are less likely to incarcerate (and convict) female than male defendants. We apply both a novel rank-order test and the recently developed disparate impact approach to assess the extent to which these courtroom gaps are driven by discrimination. We find that judges, and especially female judges, engage in gender-based discrimination. Finally, despite a short-run increase in property crime, own-gender judicial assignment improves education outcomes and decreases adult welfare use.

JEL: K40; J16; I00.
Keywords: Gender Disparities, Discrimination, Incarceration, Juvenile, Gender, Court, Long-Run Consequences.

*Authors: Ozkan Eren, Department of Economics, University of California, Riverside, Email: ozkane@ucr.edu, Randi Hjalmarsson, Department of Economics, University of Gothenburg, Email: randi.hjalmarsson@economics.gu.se, Orgul Ozturk, Department of Economics, University of South Carolina, Email: odozturk@moore.sc.edu. This study is reviewed by the University of South Carolina IRB.Hjalmarsson by the European Union (ERC, Police-Prisons-Firms, 101093345). Views and opinions expressed are however those of the authors only. Neither the European Union nor the granting authorities can be held responsible for them. We thank Robert Pettis, Eric Baxley and Ela Ender for their excellent assistance.

1 Introduction

A fair and impartial judiciary — one in which court outcomes are only affected by the evidence in the case — is a bedrock principle of the U.S. justice system. The public's perception of such impartiality informs the extent to which they trust the courts. Yet, U.S. opinion surveys reveal a perception of unequal treatment with respect to gender, race, and economic class in the criminal justice system, which has gotten worse over time (Smith et al., 2023). For instance, 63% of survey respondents thought the state courts were fair and impartial in 2014 but only 47% in 2022 (State of the State Courts, 2014 and 2023). An accumulating body of empirical evidence supports this public perception: research has found in a multitude of contexts that external case factors, such as decision maker characteristics and group identity, indeed impact case outcomes.[1] The resulting erosion of trust in the justice system is a first-order societal concern, as are the potentially costly consequences such high-stakes decisions can have for unfairly treated individuals. How concerned should society be about courtroom biases? The answer depends on both why such courtroom disparities exist and the extent to which disparate treatment translates into unequal outcomes and consequences beyond the courtroom for both the defendant and society — a topic on which there is to date no empirical evidence.

We provide the first such evidence by studying the effects of same-gender judge assignment on both high-stakes juvenile justice adjudication and incarceration decisions and long-run out-of-court socioeconomic (education, economic self-sufficiency, and crime) outcomes. Moreover, we analyze the extent to which discrimination explains the disparities between male and female judges in how they treat male and female defendants. Youths are a particularly important group to study given the critical developmental juncture at which these decisions occur; if discrimination in the courtroom affects out-of-court outcomes like education, then a lifetime of consequences can spillover onto society and even the next generation.[2] Our emphasis on gender is in part motivated by its increasing salience for juvenile crime — with a near-doubling of the female share of arrestees over the last decades.[3] However, gender is also a well-defined "group" that is sufficiently well represented amongst judges (in contrast to racial minorities) to study the potential long-run consequences.

Assignment of defendants to judges of the same "group" can affect out-of-court outcomes through mul-

[1] External factors include sequential decision-making (Chen et al., 2016; Bindler and Hjalmarsson, 2019), emotions from sports losses (Eren and Mocan, 2018), juror age, politics, and neighborhood (Anwar et al., 2014, 2019b, 2022), judge politics (Cohen and Yang, 2019), expected punishment (Bindler and Hjalmarsson, 2018), and media exposure (Lim et al, 2015; Philippe and Ouss, 2018). The contexts across these studies vary in multiple dimensions, including decision maker (judge, jury), country (U.S., France, Sweden), and even time period (historically or today).

[2] See, for example, Oreopoulos and Salvanes (2011) for a review of the nonpecuniary benefits of schooling and Hjalmarsson and Lochner (2012) for a survey of the crime impacts of education.

[3] In the U.S., 10 to 20-year-olds comprised nearly 15% of all arrests in 2019 (Office of Juvenile Justice and Delinquency Prevention). About 35 and 20 percent of property and violent crime arrests in 2019 were of females but only 20 and 10 percent, respectively, in 1980 (Puzzanchera et al., 2022).

tiple channels. First, in the case when there are group-based decision-making biases, then the most direct potential channel is via differential court decisions (e.g., acquittal or incarceration). However, the direction of such an out-of-court effect, and potentially whether there is one at all, is not clearly predicted by existing research and/or theory. On the one hand, a small body of juvenile judge instrumental variable (IV) papers (Aizer and Doyle, 2015; Eren and Mocan, 2021) find that juvenile incarceration leads to worse post-release outcomes. If the marginal offender off of which such local average treatment effects are identified generalize to the margin under which such in-group biases occur, then assignment to judges who favor members of their own group should improve long-term out-of-court outcomes. On the other hand, research (Hjalmarsson, 2009) also finds using an alternative sentencing discontinuity-based design that juvenile incarceration leads to better outcomes. The causal evidence on how incarceration impacts adult well-being is equally mixed, with a wide range of estimates across different contexts and research designs (e.g., Mueller-Smith, 2015; Garin et al., 2024).

Moreover, the extent to which disparate treatment in the courtroom has downstream effects may depend on why the disparities exist. In the case of taste-based discrimination, then Becker's (1957) classical outcomes testing framework would indeed predict unequal long-term outcomes. But if the underlying source of unequal in-group treatment is decision maker differences in their ability to evaluate in-group cases and determine who will (or will not) be harmed by incarceration, for instance, then such a long-term effect may not occur. In other words, this would be one scenario, potentially, under which the marginal offenders in the judge IV and in-group bias settings are not the same. Another channel through which in-group decision making disparities can have long-term effects is via an offender's updating of their beliefs concerning expected punishment in the future. Finally, even in the absence of unequal treatment, assignment to judges in the same group as oneself can in theory still impact individual outcomes via channels such as a role model effect. Given the time-limited nature of the juvenile-judge interaction, however, we view this channel as unlikely (in contrast to, for example, a teacher-student setting).

Two key hurdles - data and identification - exist to studying the out-of-court consequences of assignment to same-gender judges. There are three essential data requirements: (i) the ability to match court registers with administrative registers capturing long-run socioeconomic outcomes (e.g., crime, education, welfare), (ii) the observation of judicial decision maker identifiers, and (iii) data on judge characteristics (e.g., race and gender). We have assembled the first data set to meet all of these criteria. Specifically, we study more than 20,000 family court decisions from 2002 to 2012 in South Carolina. These decisions were made by 84 judges, about 30 percent of whom were female. To assign a causal interpretation to in-group judge gender assignment, one must be able to rule out that different judges see systematically different cases. Such

selection can in theory occur both within and across courthouses, as judges may select into jurisdictions with different characteristics. Conditional (on court-by-time fixed effects) random variation in South Carolina judge assignments allows us to overcome this identification challenge.

The first part of our analysis uses a generalized difference-in-differences approach to test for in-group gender disparities in judge dispositions of juvenile defendants. We regress whether a defendant is incarcerated (as opposed to being acquitted or convicted with more lenient sentences) on the interaction of defendant and judge gender, while controlling for court-by-time-fixed effects. We find significant evidence of in-group gender gaps in sentencing decisions: female judges are 2.6 percentage points (18 percent) less likely to incarcerate female than male defendants, relative to male judges. These estimates are robust to defendant and case controls, judge fixed effects, and a full set of interactions between judge (defendant) gender and other defendant (judge) characteristics. We also find substantial heterogeneity in the extent of such in-group disparities with respect to judge's professional experience in law: the in-group incarceration effect estimate is negative for the full sample of judge decisions, however, it is more than twice as large for relatively more experienced judges. Finally, although same-gender-based disparities in acquittal decisions are not apparent when pooling all judges together, more experienced judges are also more lenient (or harsher), albeit insignificantly, toward in-group (out-of-group) defendants at this margin.

We further complement our baseline results, based on the generalized difference-in-differences method, with a rank-order test that leverages the concept of stochastic dominance to establish a robust ordering across the entire distribution of outcomes. This contrasts with the typical rank-order test (e.g., Anwar and Fang 2006), which focuses on the average sentencing outcome and can therefore be influenced by a small number of decision-makers. We find substantial evidence against the rank-order independence assumption. Specifically, the gender of the juvenile defendant appears to influence the ranking of incarceration rates by judges' gender, and this violation is not merely driven by a small subset of judges. Findings like these are typically interpreted as evidence of taste-based discrimination in the literature (Anwar and Fang 2006; Alesina and La Ferrara 2014), though it does not inform us which judges (males, females, or both) are discriminatory.

The second part of our analysis turns to assessing why these in-group disparities in courtroom decisions exist. The results strongly suggest that judges discriminate on the basis of gender, with female judges being the primary drivers of this discrimination. We reach this conclusion using the disparate impact approach developed by Arnold et al. (2022) to study discrimination in bail decisions and recently applied by Baron et al. (2024) to foster care placement decisions. The South Carolina context- with conditional random assignment of cases to judges that vary in sentencing leniency – allows for the implementation of this approach.

In our context, we define disparate impact as gender differences in non-incarceration rates, conditional on a child's potential for juvenile recidivism. Basically, we compute the difference between male and female non-incarceration rates after rescaling them for their gender-specific mean risk of reoffending. Adjusting these estimates for defendant and case characteristics allows us to assess the extent to which disparate treatment is driven by discrimination on the basis of gender as opposed to other characteristics correlated with gender. Finally, we regress our judge specific measures of disparate impact on judge characteristics. Overall, we find that judges are 1.2 percentage points less likely to impose non-incarceration judgements for male than female juvenile defendants with similar risks of reoffending; in other words, holding the juvenile's propensity to recidivate constant, judges are more lenient (or harsher) towards female (male) defendants. However, it is primarily female judges underlying this pattern, as female (male) judges are 2.7 (0.7) percentage points less likely to assign a non-incarceration sentence for male versus female defendants of similar recidivism risk. The similarity of these disparate impact estimates to those from the generalized differences-in-differences design supports our use of the latter to also study the effect of same-gender judge assignment on long-run socioeconomic outcomes.

The third part of our analysis demonstrates that there are indeed out-of-court consequences of assignment to a same-gender judge. Specifically, we study the effect of same-gender judge assignment on three categories of outcomes: crime and delinquency, educational attainment and performance, and welfare receipt. Despite what we interpret as a mechanical increase in property crime when young (consistent with an incapacitation channel), we find improved education and welfare outcomes. Relative to male judges, female defendants assigned to female judges are more likely than male defendants to take the high school exit exam (10 percent), enroll in twelfth grade (23 percent), and graduate high school in four years (16 percent). Though not significant, they are also 23 percent less likely to rely on social welfare programs, as measured by participation in Temporary Assistance for Needy Families (TANF). Consistent with our results on the sentencing margin, these effects are both larger and more precisely estimated for juveniles whose cases were handled by more experienced judges. We provide several robustness checks and falsification tests supporting our research design and findings.

This paper makes significant contributions to three literatures. The first studies biases in criminal justice decision making and provides strong evidence, as noted earlier, that many factors besides the quality of evidence in a case affect verdicts and sentences. One strand of this literature, in the spirit of Akerlof and Kranton (2000), focuses on whether group identity affects criminal justice decisions. Such in-group biases have been studied in multiple U.S. contexts, both in terms of the nature of the decision and group. Studies of gender and race are done for juries and courts (Anwar et al., 2012; Boyd et al., 2015; Lim et al., 2016;

Depew et al. 2017; Flanagan, 2018; Anwar et al., 2019; Hoekstra and Street, 2021) as well as police search behavior and use of force (Anwar and Fang, 2006; Antonovics and Knight, 2009; Goncalves and Mello 2021; Hoekstra and Sloan, 2022).

The second literature studies the long-run consequences of justice system interactions for the offender. Specifically, several papers investigate causal effects of incarceration on socioeconomic outcomes, such as crime, employment, and earnings, by leveraging exogenous variation in incarceration from the quasi-random assignment of criminal cases to judges who differ in their propensity to incarcerate defendants (Kling, 2006; Aizer and Doyle, 2015; Mueller-Smith, 2015; Bhuller et at., 2020; Eren and Mocan, 2021; Norris et al., 2021; Garin et al., 2024). A number of recent papers also study other forms of criminal justice interactions, including pretrial detention and misdemeanor prosecution (Dobbie et al., 2018; Agan et al., 2023; Baron et al., 2023).

Our paper makes a number of contributions to this literature. First, only a few of the existing papers study juveniles (Aizer and Doyle, 2015; Eren and Mocan, 2021; Baron et al. 2023), finding lower educational attainment and increases in adult crime. Second, in contrast to these studies, we do not just look at education and adult crime but also adult welfare receipt. Third, our paper is the first study of the long-run consequences of justice system interactions using a design other than a judge instrument. Our finding that such in-group courtroom disparities affect an individual's long-run outcomes thus also speaks to the external validity of the harsh-judge studies (i.e., by studying a similar question with a potentially different marginal offender). Fourth, out-of-court consequences of assignment to a same-gender judge may operate through different sentencing margins, meaning we do not necessarily estimate the impact of incarceration itself, but rather the net effect of the gender interaction. This net effect is, a priori, difficult to sign.

Finally, in-group bias disparities have been documented in many non-judicial U.S. contexts, including education (Dee, 2005; Hoffman and Oreopoulos, 2009; Carrell et al., 2010), sports refereeing (Price and Wolfers, 2010; Parsons et al., 2011), and medical care and evaluations (Zeltzer, 2020; Cabral and Dillender, 2024). Though there can be downstream consequences in many of these contexts, just a handful of education papers consider this question (Fairlie et al., 2014; Gershenson et al., 2022).[4]

The remainder of the paper is organized as follows. Section 2 discusses the institutional setting. Sections 3 and 4 present the data and econometric methodology, respectively. Section 5 presents three sets of main results (and associated robustness and heterogeneity analyses): (i) the effects of in-group judge assignment on courtroom decisions, (ii) disparate impact estimates, and (iii) the effect of in-group judge assignment on

[4]See Gazal-Ayla and Sulitzeanu-Kenan (2010), Shayo and Zussman (2011); Choi et al. (2022), Ash et al. (2023), Guo et al. (2023) and Ahrsjo et al. (2024) for international evidence on group identity and criminal justice decisions. See also Bagues et al. (2017), Fisman (2017), Bar and Zussman (2019) and Ye and Yi (2023) for international evidence of in-group disparities on non-judicial contexts.

out-of-court socioeconomic outcomes. Conclusions are provided in Section 6.

2 Institutional Details

South Carolina uses a system of legislative, merit-based elections for almost all judicial positions. To vie for these positions, candidates must submit comprehensive applications to the Judicial Merit Selection Commission (JMSC), which is comprised of six members from the General Assembly (including three Senators and three Representatives) and four non-legislative members. Legislators then select judges from a pool of up to three candidates deemed qualified for judicial office by the JMSC. Judicial elections take place during a joint session of the General Assembly, with each member of the Senate and House casting one vote and the outcome being determined by a majority rule. Judges are elected to six-year terms and are subject to re-election.

The judicial system is structured into 16 judicial circuits, each comprising a combination of the state's 46 counties. The state's court of general jurisdiction is known as the circuit court. Additionally, family courts hold exclusive jurisdiction over cases involving youth under the age of 17 and matters related to family relationships.[5] Unlike a fixed assignment system, judges in South Carolina do not sit exclusively in one court or county. Instead, they travel throughout different counties within their resident circuits and to circuits outside their residential jurisdiction. The rotation system is neither uniform nor systematic and largely depends on expected caseload requirements, meaning that judge allocations to courthouses are made before the specific cases on the docket are known. Judges are assigned to courts on an approximately weekly term, although it is not uncommon for fewer days of court assignment to occur. Scheduling is under the direction and supervision of the chief justice. In our analysis sample, family court judges, on average, traveled to six counties from three judicial circuits, while counties had seven different family court judges handling juvenile defendant cases during any given year.

Having received a formal complaint, the Office of the Solicitor must decide whether to petition the juvenile case to the court. Prosecutors may choose not to do so because of a lack of sufficient evidence. If certain criteria are met, the prosecutor may also allow the case to be diverted from the juvenile justice system.[6] This means that instead of being prosecuted in the family court, the juvenile is allowed to participate in a diversion program, such as arbitration, a behavior contract, or drug court (Children's Law Center 2019). Upon successful completion of the diversion program, the charges against the child will be dismissed. Finally,

[5] South Carolina law defines "child" as a person less than 18 years old. However, in the context of delinquency and criminal acts, "child" refers to a person under age 17. Additionally, this definition excludes 16 years-old individuals charged with several felony offense categories (e.g., second-degree burglary; domestic violence of aggravated nature).

[6] These criteria include, but are not limited to, being a first-time offender, a nonviolent offender, or drug/alcohol dependent (Children's Law Center 2019).

if the prosecutor chooses to prosecute, the case moves to adjudication, which is comparable to trial in adult court. A family court judge reaches a verdict after hearing the evidence presented by the prosecution and defense attorneys.

The judge may determine that the state failed to prove its case beyond a reasonable doubt and find the child not guilty, or the judge may find the child guilty and adjudicate the child delinquent, which is analogous to a conviction in adult criminal court. If the decision is to acquit, the child is released from the court's jurisdiction. Upon a guilty verdict, the judge has to then make a disposition decision.[7] These judgements may include restitution, community service, behavioral/drug assessments and treatment, probation and incarceration (i.e., commitment to the South Carolina Department of Juvenile Justice, SCDJJ, for placement in one of its institutions). The terms of probation are also specified by the judge, and can include, for instance, regular school attendance, random drug testing, and community service. The SCDJJ operates its own school district and committed juveniles must continue their education in the agency's middle and high schools until they reach the minimum drop out age.

In general, the judge should impose the least restrictive sentencing option consistent with the circumstances of the case, the health and safety of the child, and the best interest of the community. However, the state's juvenile justice code also mandates that the sentence must hold the child accountable for their actions and prevent future violations of the law (Children's Law Center, 2019). To date, South Carolina remains a non-guideline state, meaning that judges exercise large amounts of discretion in their disposition decisions, constrained only by mandatory penalty and statutory maximums (Hester 2017).

We conclude by presenting statistics that highlight the prevalence of juvenile arrests and prosecutions in the state of South Carolina, demonstrating that South Carolina is fairly representative of the U.S. overall. Specifically, Panel A of Figure A1 in the Online Appendix displays trends in juvenile arrest rates (combining violent and property crimes) from 2002 to 2012, while Panel B shows the average proportion of juvenile referrals resulting in prosecution over the same period.

[7] After adjudicating a child delinquent, the family court judge almost always moves directly to the sentencing phase without ordering a predisposition evaluation report, which includes information about the child and their surroundings, background, and circumstances (Hester 2017). This is because judges already have access to a similar report from the initial screening process or "DJJ intake." If a dispositional hearing is held at a later date, South Carolina Code of Laws require judge who presided over the adjudicatory hearing to also preside over the disposition hearing (South Carolina Family Court Law-Rule 37).

3 Data

3.1 Data Sources

This paper matches data from multiple criminal justice and socioeconomic administrative registers. From the South Carolina Department of Juvenile Justice (SCDJJ), we obtained information on all petitioned case files from 2002 to 2012 at the charge, case, and defendant levels. The charge-level data include details about the type and severity (based on state-specific grades) of offenses and sentencing outcomes for each charge. The case-level data provide information on offense, referral and petition dates, the court where the case was tried, and the date of disposition. Finally, the defendant-level data include information on the juvenile's gender, race, and age at disposition.

The second data source is the log of judicial rotation assignments from the South Carolina Court Administration, which informs us of the presiding judge for each court and day of the calendar year.[8] We use these logs to construct a family court judge-by-court-by-date of disposition panel. We also supplement these data with manually collected information on judge attributes.[9] To link detailed case files to judges, we match administrative data from the SCDJJ to our log panel using court and disposition date. Juvenile defendants can be successfully matched to their disposition judges when cases are handled by a single judge on any day in a given court. We can match around 38 percent of all juvenile case files.[10] As discussed below, the quasi-random assignment of case files to judges minimizes concerns about sample selection bias.

To study the effects of courtroom gender interactions on outcomes beyond the courtroom, we use a common and unique individual identification number to link the matched juvenile defendant-judge data with administrative records from: (i) the South Carolina Department of Education (SCDE), (ii) the South Carolina Department of Social Services (SCDSS), and (iii) the South Carolina State Law Enforcement Division (SLED). Any juvenile defendant who does not have a match is coded as zeros for these outcome variables.

The SCDE data include information on grade enrollment and on-time high school graduation from 2000 to 2017 as well as high school exit examinations. Public school students are required to pass the high school exit exam to qualify for a diploma. The High School Assessment Program (HSAP) was first administered (in the spring of tenth grade) during the 2005-2006 academic year as the state's designated exit exam. Exit exams

[8] We scraped judicial rotations from the South Carolina Judicial Branch website (https://www.sccourts.org/calendar/).

[9] We collected demographic and other information on judges from a variety of sources, including online searches and using short bios of family court judges' available on the South Carolina Judicial Branch website.

[10] We also exclude matches in which the disposition judge handled less than 10 case files over the sample period. There are only three judges meeting this criteria. Additionally, we choose not to expand the sample of cases to those in which we can match the gender of the judge (i.e., by taking advantage of days when all judges in a courthouse are the same gender), since in such cases we would not be able to control for judge fixed effects or other observable judge characteristics.

were discontinued by the state in 2015.[11] From the SCDSS, we obtain information on enrollment in social programs through 2019. We use information on welfare receipt – both food stamps (renamed Supplemental Nutrition Assistance Program or SNAP in 2008) and Temporary Assistance for Needy Families (TANF) – to measure the long-run economic outcomes of our core sample. Finally, from SLED, we obtain adult arrest records for 2000 to 2017. For each offender file in the adult criminal system, we have basic arrestee demographic information, offense date, and type of offense committed.

3.2 Outcome Variables

Our primary sentencing outcome of interest is whether a juvenile defendant was sentenced to incarceration, defined at the case level based on the most serious conviction. Given judges can acquit juvenile defendants, we also examine the conviction margin by using an indicator variable for being acquitted. Measures of educational outcomes include: (i) HSAP test-taking, defined as a binary indicator equal to one if the child ever took the high school exit exams, (ii) twelfth grade enrollment status, and (iii) on-time high school graduation.[12] To assuage concerns about multiple hypothesis testing, we also created a schooling index by averaging the z-scores of all non-missing educational outcomes (Kling et al. 2007). We use the SCDSS data to construct two measures of economic self-sufficiency: whether the child ever received food stamps and whether the child ever received TANF as an adult. We measure criminal activity by defining separate dummies for whether the juvenile was ever arrested in early adulthood (i.e., a binary indicator equal to one if the child was arrested as an adult) and a combined measure of juvenile arrest following the first-time disposition and adult arrest (i.e., a binary indicator equal to one if the child was rearrested as a juvenile or arrested as an adult).

3.3 Sample and Descriptive Statistics

Our core analysis sample consists of 23,569 juvenile case files (from 15,591 unique children) that were prosecuted and successfully matched to the disposition judge. These files were handled by 84 judges in 46 family courthouses from 2002 to 2012. Since the juveniles were primarily born between 1988 and 1998, we can observe educational and crime outcomes up until age 29 and welfare participation up to age 31. For example, we can observe the receipt of TANF until age 31 for children born in 1988, age 30 for children

[11] South Carolina repealed the Common Core Standards in 2015. The decision to abandon exit exams was part of the state's efforts to develop its own educational standards and put in place assessments that will better identify areas of academic strength and weakness. As a result, the HSAP was replaced by end-of-course exams. Scores from these exams count towards students' final grade in gateway subjects such as Algebra, Biology and English.

[12] Approximately 12 percent of juvenile defendants took the HSAP before the disposition date. We exclude these observations from the analysis sample when the outcome of interest is HSAP test-taking. However, keeping them in the sample and imputing the dependent variable either with zero or one for these children yields point estimates that are almost identical to those presented throughout the paper.

born in 1989, and so on. Because we observe all public-school enrollments in South Carolina, concerns about attrition in the education data only arise if children leave the state, attend a private school, or are home-schooled. Endogenous mobility is arguably even less problematic for crime and welfare outcomes since attrition would only be affected by out-of-state mobility. According to data from the American Community Survey, around nine percent of adults born in South Carolina between 1988 and 1998 had left the state by 2018; the migration outflow is even lower (seven percent) when restricting these birth cohorts to have high school or lower levels of education.

Table 1 reports the case level summary statistics for our matched estimation sample (Column 1) and for those cases that are not successfully matched to their disposition judge (Column 3). Online Appendix Table A1 also shows the summary statistics by juvenile's gender. As displayed in Panel A of Table 1, among matched case files, 14 percent of juvenile defendants were incarcerated, 74 percent were convicted but not incarcerated (with more than 70 percent being placed on probation) and 12 percent were acquitted.[13] Twenty-eight percent of all case files involve female defendants. Black and white juveniles comprise 56 and 41 percent of all case files, respectively, and the average age at disposition is 15. In our sample, 21 percent of cases involve a property offense, 16 percent involve a violent offense, and 22 percent involve a school-related offense (e.g., disturbing schools; threatening a teacher/principal).[14] Finally, 32 percent of cases are associated with more than one offense.

Panel A of Table 1 also presents summary statistics for the main out-of-court outcome variables at the defendant level. The mean twelfth grade enrollment is slightly above 16 percent, while 14 percent of juvenile defendants graduated high school on-time. Around 80 (14) percent used food stamps (TANF) as an adult (age 18 and onwards) and 63 percent of the study population was arrested in early adulthood.[15]

Panel B summarizes the demographic characteristics of the family court judges in the matched sample. Around 26 percent of juvenile cases were handled by female judges, while white judges handled almost 95 percent of all cases. The average judge age at disposition is 53 and 86 percent of juvenile cases were presided over by judges who graduated from a college in South Carolina. Family court judges have, on average, 22 years of professional legal experience, measured by the number of years between graduation from law school and their first appearance in the data.

We next turn to how representative the matched sample is of the universe of South Carolina juvenile

[13] Mandatory school attendance comprises a large fraction (aorund 67 percent) of the judgments within the subcategory of convictions without probation. The remaining judgements are a very heterogeneous group, ranging from house arrest to community service to behavioral/drug assessments.

[14] The remaining 40 percent involve offenses in the "other" charges category. This is a very heterogeneous group and includes offenses ranging from ungovernable to unlawful carrying of a pistol.

[15] The food stamps program has a significantly larger base of participation than TANF. It provides a steady stream of benefits to income and asset-eligible households, as well as able-bodied adults without dependents. TANF, on the other hand, provides financial assistance to low-income households with dependent children.

court cases. Online Appendix Figure A2 displays the number and share of case files that were successfully matched in each of South Carolina's 16 judicial districts. This share is less than 20 percent in three districts and between 40 and 100 percent in eight districts; there is a significant number of cases from all districts. Panel C of Table 1 looks closer at the underlying county characteristics for the matched and unmatched samples.[16] Cases from less populated counties are over-represented in the matched sample. This is not surprising given that matching could only be done on courthouse-days with just one family court judge. Consistent with this, Panel C shows that defendants in the matched sample are tried in counties with on average about seven family court judges while this statistic is nearly double in the unmatched sample.

But, does this non-representativeness in terms of county characteristics (e.g., ruralness and population size) impact the representativeness of the sample in terms of juvenile defendant/case characteristics and both sentencing and long-run socioeconomic outcomes? Columns 1 and 3 of Table 1 demonstrate that the matched and unmatched samples are surprisingly similar. To shed further light on the extent to which there are sample selection issues, Online Appendix Table A2 regresses judge missingness (i.e., whether a case is matched to a judge) on defendant and case characteristics (Column 1) and case outcomes (Column 2). These observable characteristics have little predictive power for either the full sample or the subsamples of male and female defendants (Columns 3 and 4); the p-values for joint significance range from 0.34 to 0.57. These results support the conclusion that failing to match cases to disposition judges is as good as random.

4 Empirical Methodology

4.1 Difference-in-Differences Approach

To evaluate juvenile defendant-judge gender interaction effects on judicial dispositions, we use a generalized difference-in-differences approach and estimate the following equation:

$$JD_{icjt} = \beta_0 + \beta_1 Female_Def * Female_Judge_{icjt} + X'_{icjt}\beta_2 + \lambda_j + \gamma_{ct} + \epsilon_{icjt} \quad (1)$$

where JD_{icjt} is the disposition outcome for juvenile defendant i in case c, as determined by judge j at time t. For example, this variable takes the value of one if the juvenile defendant is adjudicated guilty and incarcerated and zero if the child was acquitted or adjudicated guilty with a sentence of non-incarceration (e.g., placed on probation). X'_{icjt} is a vector of observed case and juvenile defendant characteristics (i.e., indicators for gender and race, offense severity, offense type fixed effects, year of birth fixed effects, number

[16] The descriptive statistics in Panels B and C, which are calculated over all juvenile case files, do not deviate in a meaningful way when the data is collapsed at the judge or county level.

of charges at disposition, and number of times a juvenile defendant is observed in the data). λ_j and γ_{ct} stand for judge and court-by-time fixed effects, respectively and ϵ_{ijct} is the error term.

Because case types may vary by court, year-month, and day of the week, we incorporate court-by-year-month and court-by-day-of-week fixed effects into our specifications. The inclusion of court-by-time fixed effects (γ_{ct}) safeguards against potential case selection and ensures that family court judges preside over similar case files, a condition that must hold for causal identification. Furthermore, controlling for judge fixed effects addresses the possibility of non-random sorting of judges across jurisdictions. For example, it is plausible that female judges, with tendencies toward harsher punishments or differential gender attitudes, may systematically choose to serve in judicial circuits (and thereby counties) where gender (of either the defendant or the victim) is a particularly salient case characteristic, such as judicial districts with higher prevalence of sex crimes. In this case, the difference-in-differences estimates without judge fixed effects would capture the combined effects of both within-judge behavior and differential sorting. Conditioning on judge fixed effects in equation (1) avoids such non-random sorting potentially contaminating the estimated gender interaction effects. Robust standard errors are two-way clustered at the judge and court level.

In this specification, the coefficient β_1 reflects the average change in incarceration (or conviction) rates for female defendants, relative to males, from the quasi-random assignment of case files to female versus male judges. Put differently, we estimate whether gender disparities in sentencing outcomes are different between female and male judges. This measure is necessarily a relative one and we cannot sort out whether a non-zero effect means female judges impose more lenient (harsher) dispositions on female (male) juvenile defendants, or whether male judges impose more lenient (harsher) dispositions on males (females). Answering this question in this framework is difficult because it requires the establishment of a (non-discriminatory) benchmark, which is not easy due to the lack of sentencing guidelines in South Carolina.

In addition to examining judicial decisions, we also use this research design to analyze the courtroom gender interaction effects on long-run educational, economic self-sufficiency, and criminal justice outcomes.[17]

The key identifying assumption in this empirical framework is that the variation in judge gender is as good as random when conditioning on court-by-time fixed effects. Table 2 empirically assesses the validity of our research design. The first column uses a linear probability model to test whether case and defendant characteristics are predictive of incarceration, after controlling for court-by-time fixed effects. Each cell represents a separate regression. We find that the decision to incarcerate a juvenile defendant is highly correlated with all observable covariates. The coefficient estimates are also jointly significant from a

[17] Since youth may have multiple encounters with the juvenile justice system and thus be prosecuted more than once, some children are observed multiple times in our analysis sample. Consequently, all of our long-run specifications are estimated using probability weights, where the weight is the inverse of the number of times a juvenile is observed in the sample. These models do not control for number of times a juvenile defendant is observed in the data.

single regression using all juvenile/case characteristics (p-value=0.00). As shown in Column 2, observable characteristics are also highly correlated with the decision to acquit a juvenile defendant. Columns 3-5 of Table 2 present the point estimates using judge characteristics (gender, race/ethnicity, and age) as dependent variables. None of the nine observable characteristics are significantly related (at even the 10% level) to judge gender (Column 3); the p-value for joint significance is 0.54. These results do not change in a meaningful way when judge gender is replaced with a binary indicator for judge's minority status and age (Columns 4 and 5).

We further complement these randomization tests by analyzing whether judge leniency is correlated with observable characteristics (Column 6). We measure leniency by the leave-one-out average non-incarceration rate among all other juvenile defendants assigned to a judge. Specifically, we construct the leave-one-out measure using the residuals from a regression of the non-incarceration indicator on court-by-time fixed effects.[18] Reassuringly, the coefficient estimates are small in magnitude and not statistically different from zero (p-value=0.73). The p-values for joint significance from analogous regressions for female and male juvenile defendant samples are 0.29 and 0.90, respectively. Online Figure A3 also displays the distribution of the residualized judge leniency, which ranges from -0.138 to 0.054 with a standard deviation of 0.022.

Finally, the difference-in-differences estimates can be confounded if more male or female judges are not responding to defendant gender, but to some other defendant or case characteristics that are systematically correlated with gender. For example, if female judges are more likely than males to incarcerate defendants charged with violent offenses, and if there is a higher proportion of male defendants charged with these offenses, then our estimates could be biased. The results may also be driven by how juveniles respond to judges. It is possible that defendants might be responding to some other characteristics of the judge that are correlated with gender. To address these concerns, we demonstrate the robustness of the estimates to including controls that interact both defendant and judge gender with juvenile characteristics (e.g., race and offense type) and judge characteristics (e.g., race and experience).

4.2 Disparate Impact Approach

To investigate the source of differential gender disparities in sentencing outcomes by judge's gender, we follow the methodology developed in Arnold et al. (2022) and estimate what they refer to as disparate impact. In our context, disparate impact can be defined as the gender differences in non-incarceration rates, conditional on a child's potential for juvenile recidivism. At its core, the approach involves rescaling female and male non-incarceration rates based on their gender-specific mean risk of reoffending and then computing the difference

[18] To allow for meaningful variation in judge leniency in a given cell, court-by-time fixed effects for leniency (and for disparate impacts) specifications include court-by-year-by-quarter fixed effects.

between these adjusted rates. The resulting estimate—disparate impact—captures both direct gender-based discrimination (whether through taste-based or statistical discrimination) and indirect discrimination via non-gender characteristics that are correlated with gender.[19] Adjusting for defendant and case characteristics in the estimation further isolates direct discrimination on the basis of gender (Baron et al. 2024). Given that one of the primary objectives of family court judges is to prevent a child's future violation of the law (Children's Law Center, 2019), our definition of disparate impact is well-founded.

The main obstacle in estimating disparate impact is the selective observability of juvenile recidivism. While the risk of a child's reoffending is directly observed for non-incarcerated juveniles, it is either unobserved or only partially observed for those placed in confinement due to incapacitation. To address this challenge, we exploit the conditional random assignment of cases to judges, and the fact that judges vary in their leniency. These institutional features of the South Carolina juvenile justice system allow us to estimate gender-specific juvenile recidivism potential by examining the subsequent outcomes of cases handled by a hypothetical supremely lenient judge who avoids an incarceration disposition for all juvenile defendants. In the absence of such a lenient judge, one can estimate the average juvenile recidivism risk for female and male defendants by extrapolating the trend observed among different judges, as if such a hypothetical judge existed.

We now formally define disparate impact. Consider a binary variable $D_{ij} \in \{0, 1\}$ for juvenile defendant i disposed by judge j, where $D_{ij} = 1$ denotes a non-incarceration sentencing outcome. $Y_i^* \in \{0, 1\}$ represents the juvenile's recidivism potential, indicating whether defendant i would reoffend if acquitted or adjudicated guilty with a sentence of non-incarceration (e.g., probation). The observed recidivism outcome, $Y_i \in \{0, 1\}$, equals Y_i^* when the juvenile is not incarcerated ($D_{ij} = 1$) and is either unobserved or only partially observed when the juvenile is incarcerated. The statewide average time served for incarcerated juveniles is slightly above 200 days (South Carolina Department of Juvenile Justice Data Resource Guide, 2017). Accordingly, in our paper, we define juvenile recidivism as the incidence of rearrest within six months following the disposition date. Our findings remain robust to alternative definitions of the recidivism timeframe, such as rearrest within three months or one year of the disposition date.

The disparate impact for each judge is defined as the average difference in non-incarceration rates between female and male defendants with the same juvenile recidivism potential. Specifically, let $G_i \in \{f, m\}$ denote the gender of the child and define

$$\Delta_{j0} = E[D_{ij}|G_i = m, Y_i^* = 0] - E[D_{ij}|G_i = f, Y_i^* = 0] \tag{2}$$

[19] Note that indirect discrimination is similar to judges responding to factors that are systematically associated with gender, as discussed in the difference-in-differences approach.

as the non-incarceration disparity between male and female defendants without any juvenile recidivism potential, and

$$\Delta_{j1} = E[D_{ij}|G_i = m, Y_i^* = 1] - E[D_{ij}|G_i = f, Y_i^* = 1] \tag{3}$$

as the non-incarceration disparity between male and female defendants with juvenile recidivism potential. The average disparate impact of judge j is then given by the weighted average of equations (2) and (3):

$$\Delta_j = \{1 - \Pr(Y_i^*)\}\Delta_{j0} + \Pr(Y_i^*)\Delta_{j1}$$

where $\Pr(Y_i^*) = E[Y_i^*] = \bar{\mu}$. Following this, the average disparate impact of judge j can be written as:

$$\Delta_j = (1 - \bar{\mu})\Delta_{j0} + \bar{\mu}\Delta_{j1} \tag{4}$$

Using equation (4), we infer that judge j discriminates against male (female) defendants when $\Delta_j < 0$ ($\Delta_j > 0$) and does not discriminate when $\Delta_j = 0$. As noted above, the main challenge in estimating the disparate impact is the selective observability of juvenile recidivism. Following Arnold et al. (2022) and using the law of iterated expectations, we can show that

$$E[D_{ij}|G_i = g, Y_i^* = 0] = \frac{E[D_{ij}(1 - Y_i)|G_i = g]}{1 - \mu_g} \tag{5}$$

where $\mu_g = E[Y_i^*|G_i = g]$ represents the gender-specific average juvenile recidivism risk. Under the assumption of random assignment of cases to judges, we obtain the following equality:

$$\begin{aligned} \delta_{jg0} &= E[D_{ij}|G_i = g, Y_i^* = 0] \\ &= \frac{E[D_{ij}(1 - Y_i)|G_i = g]}{1 - \mu_g} = \frac{E[D_{ij}(1 - Y_i)|G_i = g, Z_{ij} = 1]}{1 - \mu_g} \end{aligned} \tag{6}$$

where Z_{ij} is a binary variable indicating that defendant i was assigned to judge j. Similarly,

$$\begin{aligned} \delta_{jg1} &= E[D_{ij}|G_i = g, Y_i^* = 1] \\ &= \frac{E[D_{ij}Y_i|G_i = g]}{\mu_g} = \frac{E[D_{ij}Y_i|G_i = g, Z_{ij} = 1]}{\mu_g} \end{aligned} \tag{7}$$

Substituting equations (6) and (7) into the expression for the average disparate impact of judge j from

equation (4) and rearranging the terms yields

$$\Delta_j = \bar{\mu}\{\delta_{jm1} - \delta_{jf1}\} + (1-\bar{\mu})\{\delta_{jm0} - \delta_{jf0}\} \quad (8)$$

$$= E[\Omega_i D_i | G_i = m, Z_{ij} = 1] - E[\Omega_i D_i | G_i = f, Z_{ij} = 1] \quad (9)$$

where $\Omega_i = \dfrac{\bar{\mu} Y_i}{\mu_{G_i}} - \dfrac{(1-\bar{\mu})(1-Y_i)}{1-\mu_{G_i}}$. Note that Ω_i is a function of observed juvenile recidivism outcome (Y_i), unobserved average gender-specific recidivism risk (μ_g), and overall average risk ($\bar{\mu}$), which is given by $\bar{\mu} = \Pr(G_i = m)\mu_m + \Pr(G_i = f)\mu_f$. Since Y_i and D_{ij} are observed in the data, estimating the disparate impact defined in equation (9) only requires knowledge of μ_g for the full population of female and male juvenile defendants.

In doing so, we leverage the randomization of case files to judges and the fact that juvenile defendants assigned to each judge have the same underlying recidivism risk on average. To further develop the intuition for this approach, consider the existence of a supremely lenient judge j^* who does not incarcerate any juvenile defendants, regardless of their gender or recidivism potential. In a setting with as-good-as random assignment, this implies

$$\mu_g = E[Y_i^* | G_i = g] = E[Y_i^* | G_i = g, Z_{ij} = 1] = E[Y_i^* | G_i = g, Z_{ij^*} = 1]$$

and since $D_{ij^*} = 1$ for all juvenile defendants whose cases were handled by this supremely lenient judge, the observed gender-specific juvenile recidivism rate would be very close to the unobserved average gender-specific recidivism risk in the full population. That is,

$$\mu_g = E[Y_i^* | G_i = g, Z_{ij^*} = 1] = E[Y_i | D_i = 1, G_i = g, Z_{ij^*} = 1] \quad (10)$$

In practice, following Arnold et al. (2022), we estimate equation (10) using statistical extrapolations. More precisely, we run a regression of $E[Y_i | D_i = 1, G_i = g, Z_{ij} = 1]$ on $E[D_i = 1 | G_i = g, Z_{ij} = 1]$ and evaluate $E[Y_i | D_i = 1, G_i = g, Z_{ij} = 1]$ at $E[D_i = 1 | G_i = g, Z_{ij} = 1] = 1$ to obtain μ_g. This approach is warranted for two reasons. First, no judge in the data exclusively imposes non-incarceration sentences. Second, even if a supremely lenient judge were present, we aim to minimize sampling error that may arise due to the finite number of cases handled by such a judge.

Finally, we use judge-specific disparate impacts, Δ_j, to identify whether gender discrimination in sentencing decisions stems from the behavior of female judges, male judges, or both.

5 Results

5.1 Juvenile Defendant-Judge Interactions and Sentencing Outcomes

We report our baseline results for sentencing outcomes in Table 3. The estimated impact in each column comes from a separate regression. Column 1 presents the juvenile defendant-judge female interaction effect on incarceration rates when only controlling for court-by-time fixed effects. Relative to male judges, female judges are 2.4 percentage points less likely to incarcerate female than male defendants. This is roughly equivalent to the baseline sentencing gap between female and male judges and 17 percent of the mean incarceration rate. The estimated coefficient on the female interaction term is robust to including defendant and case characteristics in Column 2, which is consistent with quasi-random assignment of case files to judges. Column 3 includes judge fixed effects to account for the potential non-random sorting of female and male judges across counties. The difference-in-differences estimate remains stable and significant.

Thus far, we have found that there is a strong defendant-judge gender interaction effect in the likelihood of incarceration, as opposed to probation and other judgements or acquittal. But, is this effect driven by the sentencing and/or conviction decisions? Column 4 of Table 3 answers this question by using an indicator for a verdict of acquittal as the dependent variable in equation (1) as opposed to incarceration. The gender interaction effect on the likelihood of being convicted is small in magnitude and statistically insignificant although, as discussed in Section 5.4, this finding masks important heterogeneity (with regards to judge experience) in the estimated impact of being assigned to a same-gender judge on the acquittal margin. Finally, as an additional validation check of our research design, we use the predicted value of juvenile incarceration – obtained from a regression of an indicator for incarceration on (pre-determined) observable juvenile characteristics – as the outcome of interest in the last column of Table 3. Reassuringly, the point estimate is both economically and statistically insignificant.[20]

To put the magnitude of the estimate for incarceration in perspective, we first compare the female interaction effect of 2.6 percentage points to the unconditional juvenile female-male gap in incarceration rates and find that the effect size is roughly equivalent to 45 percent of the gender gap (Online Appendix Table A1). Next, we compare the finding of an 18 percent reduction (relative to the baseline mean) from our preferred specification in Column 3 to previous work examining the impact of extra-judicial factors. For example, Abrams et al. (2012) find that the gap in incarceration rates between white and black defendants increases by more than 20 percent when moving from a judge at the 25th percentile to one at the 75th percentile in the racial gap distribution. Aizer and Doyle (2015) show that being assigned to a judge with a

[20] The estimated gender interaction effect is -0.003 (s.e.=0.002) when using the predicted value of acquittal as the dependent variable.

10 percentage points higher incarceration rate increases the juvenile defendant's probability of incarceration by about 11 percent. The juvenile defendant-judge gender interaction effect identified in this paper is comparable in magnitude to the effect of a change equivalent to the interquartile range of the racial gap in judge fixed effects. It is also roughly equal to an increase of being assigned to a judge with a 16 percentage points higher incarceration rate.

5.2 Individual Judges Effects and Rank Order Tests

To present further evidence on the nature of juvenile defendant-judge gender interactions, we first estimate a judge random effect model. Specifically, we regress the incarceration indicator on court-by-time fixed effects, along with case, juvenile, and observable judge characteristics, and retain the residuals. We then use these residuals to estimate a random effects model, obtaining empirical Bayes estimates for each individual judge by defendant gender.[21] The conditional randomization of case files to judges allows us to causally identify individual judge effects. Panel A of Figure 1 displays density plots of the individual judge effect estimates comparing male and female judges for female juvenile defendants. There are substantially more female judges in the left tail of the distribution. Panel B presents the corresponding judge effect estimates for male defendants – a noticeable rightward shift is evident in the distribution of female judges compared to male judges.

We next utilize stochastic dominance tests, which offer social welfare comparisons over a large class of preferences, to establish robust ordering over the entire distribution. Such comparisons may provide an important advantage over the usual rank-order test (Anwar and Fang, 2006), which centers on the average outcomes and thus can be impacted by only a few decision-makers. Suppose S_f^F and S_f^M denote the empirical Bayes estimates for female and male judges, respectively; these map to the individual judge effect estimates for female defendants (f). Let $F_f^F(s) = \Pr(S_f^F < s)$ represent the cumulative density function (CDF) of S_f^F; $F_f^M(s)$ is similarly defined for S_f^M. Under this notation, S_f^M first-order stochastically dominates (FSD) S_f^F if

$$F_f^M(s) \leq F_f^F(s) \quad \forall s \in S_f \text{ with strict inequality for some } s \tag{11}$$

where S_f denotes the union support of S_f^F and S_f^M. One can similarly formulate the FSD condition for male and female judges using individual judge effects for male defendants (m). Specifically, S_m^F FSD S_m^M can be written as

$$F_m^F(s) \leq F_m^M(s) \quad \forall s \in S_m \text{ with strict inequality for some } s. \tag{12}$$

[21] In practice, we use the Stata package *mixed*.

To test for FSD in equations (11) and (12), we utilize the Kolmogorov-Smirnov test criteria and the resampling procedure described in Linton et al. (2005). Under this resampling scheme, one can infer FSD to a desirable degree of confidence if the p-values of the test statistics are large, say 0.90 or higher. Appendix B provides the details of consistent testing procedure for stochastic dominance. Assessing the SD tests, we find that the p-value associated with S_f^M FSD S_f^F is 0.92, the p-value for S_m^F FSD S_m^M is 0.97.[22]

Taken together, these results are consistent with two potential mechanisms. The first obviously pertains to same-gender (in-group) bias on the part of judges. In our context, judicial bias implies that the gender of the juvenile defendant should affect the ranking of incarceration rates over judges' gender. This is exactly what we find. Failing to preserve FSD ranking of incarceration rates by judge's gender (i.e, S_f^M FSD S_f^F, while S_m^F FSD S_m^M) may lend support for taste-based discrimination (Anwar and Fang 2006; Alesina and La Ferrara 2014). Second, it is possible that juvenile defendants respond differently to opposite-gender judges even if judges treated all juveniles the same way. For example, female defendants may exhibit greater cooperation in the courtroom when facing a judge of the same-gender, and this, in turn, could influence sentencing decisions. Although we cannot completely rule out this alternative explanation, we believe it is unlikely for several reasons. First, the disparate impact approach in the next section relies on a comparison between female and male juvenile defendants with the same risk of reoffending. There is no a priori reason to expect defendants with identical risk levels to behave differently in the courtroom. Second, such a response does not align well with various heterogeneity analyses based on judge characteristics either (discussed below). Finally, judges determine courtroom outcomes after gathering information about the child's background, surroundings, and circumstances from predisposition and DJJ intake reports. Thus, differences in sentencing outcomes are ultimately driven by judicial decision-making.

5.3 Disparate Impact Results

We present the disparate impact estimates in Table 4 using both local linear and linear extrapolations, measuring risk of reoffending as the incidence of rearrest within six months following the initial disposition date. To identify direct discrimination (either through taste-based or statistical discrimination) on the basis of gender, we adjust the estimates for defendant and case characteristics. Robust standard errors are two-way clustered at the judge and court level and are obtained via bootstrapping procedure. As shown in the first column, the average recidivism rate within six months of the judicial disposition is 0.109 (s.e.=0.010) in the population of female children and 0.139 (s.e.=0.013) in the population of male children. That is, absent

[22] The point estimate from regressing an incarceration indicator on a female judge dummy for the sample of female juvenile defendants, while controlling for case and juvenile characteristics and court-by-time fixed effects, equals -0.014 (s.e.=0.028). The same point estimate using the sample of male juvenile defendants is 0.020 (s.e.=0.015). These average effects also point out the violation of the rank independence condition.

of incarceration, male defendants would be 3 percentage points more likely to reoffend as juveniles. Panel B shows that the case-weighted average disparate impact among all judges is -0.012 (s.e.=0.006), indicating that judges are 1.2 percentage points less likely to impose non-incarceration judgements for a male juvenile defendant than a female defendant with similar risk of reoffending. The results are not sensitive to alternative extrapolations (Column 2) and different timeframes for juvenile recidivism (Online Appendix Table A3).

However, this average impact masks significant heterogeneity, as shown in Table 5, where we regress judge-level disparate impact estimates on judge characteristics. Male judges are estimated to have an average Δ_j of -0.7 percentage points (control mean in the bottom row), while female judges have an average Δ_j of -0.025 (Column 1). This means that female judges are 2.5 percentage points are more likely to incarcerate a male juvenile defendant than a female juvenile defendant with similar recidivism potential. Importantly, gender discrimination by female judges relative to male judges obtained from the disparate impact approach (an average of 2 percentage points) is very similar to the juvenile defendant-judge gender interaction effect estimate obtained in Table 3.[23] Albeit insignificant, as shown in Column 2, it is also worth noting that we find higher levels of disparate impact among judges with above average professional legal experience, which is 22 years, and for minority judges (Column 4). Given that we lose around 20 percent of the judge sample because of missing information on judge race and that there are only a handful of minority judges, we take this result as only suggestive.

Finally, in our disparate impact analysis, we implicitly assume that judge assignment systematically affects juvenile recidivism potential only through the incarceration decision (i.e., the exclusion restriction). However, this assumption may not hold if, for example, probation terms—such as random drug testing or school attendance—meaningfully influence the risk of reoffending. To address this possibility, following Arnold et al. (2022) and Baron et al. (2024), we construct exclusion-free bounds using aggregate (state-level) rates of non-incarceration and juvenile recidivism.[24] For female judges, we obtain bounds on average disparate impact of [-0.017, -0.023]. An analogous bound on overall average disparate impact is [0.002, -0.014]. These are similar to our baseline estimates based on conditional random assignment and exclusion restriction assumptions.

Taken together, the results on judicial decision-making suggest that judges discriminate based on gender, with female judges being the primary drivers of this discrimination. Additionally, the similarity in estimated effects between the generalized difference-in-differences and disparate impact approaches reinforces the valid-

[23] The estimated gender interaction effect on incarceration is -0.018 (s.e.=0.011) when we limit the analysis sample to only include initial dispositions.

[24] The lower bound is obtained by assuming that no juvenile defendant placed in confinement would reoffend, i.e., $\mu_g = E[Y_i^* D_i | G_i = g]$. The upper bound is obtained by assuming that all juvenile defendants placed in confinement would reoffend, i.e., $\mu_g = 1 - E[D_i | G_i = g] + E[Y_i^* D_i | G_i = g]$.

ity of our reduced-form specifications to investigate the long-run effects of gender discrimination in judicial decision-making.

5.4 Juvenile Defendant-Judge Gender Interactions and Longer-Run Outcomes

This section considers the potential out-of-court consequences of these in court juvenile defendant-judge gender interactions. The expected sign of these effects is a priori ambiguous given the multiple channels highlighted in the introduction through which juvenile out-of-court behaviors and long run outcomes can be affected. These include, for instance: (i) direct short-run incapacitation effects of differential sanctions, (ii) long-run (rehabilitative or criminalizing) effects of the sanctions, and (iii) belief updating about expected punishment in the future. We re-estimate our baseline specification (Column 3 of Table 3) when using three categories of out-of-court outcomes in the short and long-run as the dependent variable: delinquency and criminal behavior, education, and welfare receipt.

We begin in Figure 2 with the crime outcomes.[25] Panel A considers the sample of juveniles receiving their first disposition and plots the coefficient on the juvenile defendant-judge gender interactions for a series of regressions where the dependent variable is whether the juvenile was arrested by age A (17, 18, 19, 20, 21, all ages) for offense Y (i.e., a combined measure of criminal involvement including juvenile rearrest and adult arrest). We consider the following broad offense categories: any (circles), drug (diamonds), property (triangles), violent (squares), and other (xs). By age 17, we see a significant increase in property offenses, but no effects on the other offense categories. Such an increase is consistent with an incapacitation effect – or the lack thereof – of the decreased likelihood of incarceration. The property offense point estimate remains fairly stable in size (and gets smaller relative to the mean) as the sample gets older; there does not appear to be an additional increase in the cumulative property crime effect. The point estimates associated with the other crime categories are insignificant and they all turn negative as juveniles age. For example, a two-sided test for any crime including all ages cannot rule out modest reductions of about 6.6 percentage points, or 10 percent over the baseline. Panel B of Figure 2 re-estimates the results for the full sample, i.e., not just first-time dispositions, when the dependent variable equals one if the individual was arrested as an adult in each offense category. A significant effect is again seen for property crime, which is likely still driven by "incapacitation" effects and all other coefficient estimates are close to zero.

The results for four education outcomes – took the high school exit exam (spring of tenth grade and beyond), enrolled in twelfth grade, graduated from high school in four years, and the composite schooling index – are presented in Columns 1-4 of Table 6, respectively. The coefficients on the juvenile defendant-

[25] See Online Appendix Table A4 for juvenile and adult arrest statistics by offense category.

judge gender interaction are positive and significant in all four specifications. Relative to male judges, female defendants assigned to a female judge are 4.2 percentage points (or about 10 percent relative to the mean) more likely to take the high school exam than male defendants. Enrollment in twelfth grade increases by 3.9 percentage points (23 percent) and the chance of graduating high school in four years increases by 2.2 percentage points (almost 16 percent).

The last two columns of Table 6 present results for our two measures of welfare use – whether the child ever received TANF (Column 5) or food stamps/SNAP (Column 6) as an adult. There is a null (albeit imprecise) effect associated with the use of food stamps but a large reduction in TANF receipt; though not quite achieving significance at traditional levels, we do see that the likelihood of TANF receipt as an adult decreases by 3.2 percentage points or 23 percent.

Recall that we have provided suggestive evidence of larger gender disparities in judges' disposition decisions as their professional experience increases (Column 2 of Table 5). To further examine short- and long-run consequences of this potential heterogeneity, we split the sample by judges' average total number of years of experience. Prior to discussing these results, it is important to emphasize that the coefficients we report reflect how being assigned to a more or less experienced judge is causally linked to a child's outcomes. However, these estimates may not fully capture the causal effect of a judge's professional experience on the juvenile, as it can be correlated with a judge's unobservable traits. Online Appendix Table A5 shows observable judge characteristics by experience. Perhaps not surprisingly, we find that judges with more experience are also older at the time of disposition.

The first column of Table 7 presents the results for the incarceration decision. The judge-juvenile gender interaction term is negative for both experienced (Panel A) and less experienced (Panel B) judges. However, the estimated effect is much larger in Panel A. Notably, albeit falling short of statistical significance, we also observe a non-trivial gender interaction effect on the acquittal margin (Column 2 of Panel A): the gender disparity in conviction rates by judge's gender is approximately 24 percent (relative to the control mean). This finding suggests that while incarceration is an important aspect of potential behavioral differences on the part of the judges, differential treatment in other sentencing margins —such as those at the conviction or sentence length— may impact long-run outcomes. Therefore, when studying out-of-court outcomes in equation (1), we do not necessarily estimate long-run consequences of incarceration decision per se, but rather the *net* gender interaction effect.

As shown in the remaining columns of Table 5, we find more favorable long-run outcomes for juveniles whose case files were handled by more experienced own-gender judges. The coefficient estimates for schooling and TANF receipt are both statistically significant in Panel A, while it is positive but insignificant (almost

equal to zero) for the schooling index (TANF) in Panel B. The confidence interval in the last column of Panel A is not precise enough to rule out non-negligible reductions in adult criminal involvement among juveniles assigned to a same-gender judge, with a two-sided test only able to rule out around a 11 percent reduction. The results of this heterogeneity analysis are not sensitive to different cutoffs for years of experience (e.g., 20 years) when partitioning the sample.[26]

In summary, there are indeed long-run consequences of juvenile assignment to same-gender judges in the courtroom. Despite a mechanical increase in property crime while young, there are improved education and welfare outcomes and possibly lower overall criminal involvement. Our reduced form analyses do not explicitly attempt to establish a causal relationship between sentencing and out-of-court outcomes for juveniles whose sentencing decisions are influenced by own-gender judicial assignments (i.e., compliers). We refrain from such an extension because, unlike judge IV designs, the exclusion restriction assumption for instrument validity appears less tenable in our context.

5.5 Robustness Checks

We next implement a series of sensitivity checks to examine the robustness of our main difference-in-differences results. We highlight the most important ones in Table 8 (Panel A: Incarceration, Panel B: Schooling Index, Panel C: TANF receipt) and present a series of additional tests in Online Appendix Tables A6-A10 and Figures A4-A5.

First, as noted above, these estimates could be biased if it is not judge or defendant gender per se that matters, but rather other characteristics correlated with gender. To probe these concerns, we control for interactions between judge gender and juvenile/case characteristics (race, severity of the offense, offense fixed effects and number of charges at disposition) and interactions between defendant gender and observable judge characteristics (race, years of experience, and an indicator for whether the judge graduated from a college in South Carolina). We lose around 25 percent of the original sample due to missing judge information. Column 1 of Table 8 shows that estimates of the own-gender effects remain large and statistically significant (and are now significant at the 10% level for TANF).

Second, instead of court-by-year-month fixed effects, we demonstrate that the results are robust to alternatively conditioning on court-by-year-week fixed effects and court-by-year-week-day fixed effects in

[26] We also examine the length of time individuals receive TANF assistance by collapsing the data into a single observation per person and constructing a variable indicating the proportion of observed years (18 and above) during which individuals received benefits. This is a continuous measure that takes on values between zero and one, where zero indicates that the individual never received TANF assistance in any of the observed periods, and one indicates that the individual received benefits in all periods. The coefficient estimate for TANF from this exercise is -0.014 (s.e. = 0.008) for juvenile defendants whose case files were handled by more experienced judges, while the impact from less experienced judges' sample is 0.004 (s.e. = 0.012).

Columns 2 and 3 of Table 8, respectively.[27] For example, the results also change very little if we (i) exclude court-by-day-of-week fixed effects from the baseline model and (ii) further replace court-by-year-month fixed effects with court-by-year-by-quarter fixed effects. The coefficients on the juvenile defendant-judge gender interaction from a specification that only conditions on court-by-year-by-quarter fixed effects are as follows: -0.020 (s.e. = 0.011) for incarceration, 0.086 (s.e. = 0.032) for the schooling index, and -0.031 (s.e. = 0.025) for TANF receipt. Third, Column 4 of Table 8 demonstrates the robustness of the results to controlling for detailed offense fixed effects (there are more than 350 unique juvenile offense types) as opposed to the previously defined broad offense categories. Finally, to address concerns about a potential mechanical relationship between mandatory school attendance orders and juvenile outcomes, we re-ran the regressions, excluding all non-incarceration convictions involving mandatory attendance. As shown in the last column, the findings remain stable.[28]

An additional concern that may arise is whether there is any judge shopping. Randomization and balancing checks already suggest that this does not occur. We further tackle the concern, however, of whether prosecutors or defense attorneys have the opportunity to influence which judge is going to preside over the case by dividing the analysis sample based on the number of days since the initial referral to the juvenile justice system or the number of days since prosecution. These results are presented in Online Appendix Table A6. Any potential judge shopping should arguably be less viable when there is a short gap between the referral/prosecution and disposition dates. The lack of any pronounced heterogeneity across different time windows is consistent with judge shopping not being a salient issue.

Online Appendix Table A7 demonstrates that the results are robust to restricting the data to those with only one disposition as well as to not weighting the estimates for long-run outcomes. Online Appendix Table A8 rotates through dropping each of the five broad juvenile offense categories and re-estimates equation (1) for both short- and long-run outcomes, an exercise that also sheds light on the heterogeneity by offense types. The estimated coefficients are qualitatively similar (though precision is sometimes lost due to large reductions in sample sizes). We also assess robustness to the level of clustering and calculate standard errors by clustering at the (i) court level, (ii) child level, and (iii) court-by-year-month level. The statistical inference remains intact and even becomes more precise for the TANF specification (Online Appendix Table A9).

Finally, we conducted two placebo tests. In the first, we used the referral date, rather than the date of disposition, to link case files to judges. If our results are driven by unobservables that are correlated with

[27] The judge fixed effects are dropped in the specification that controls for court-by-year-week-day fixed effects because of multicollinearity.

[28] More directly, regressions with mandatory school attendance as the outcome variable show that differential probation terms are not underlying the education effects; the point estimate on the gender interaction terms is -0.005 (s.e. = 0.012).

judge gender that affect sentencing, then we would expect to see a significant spurious correlation between the female interaction term and sentencing and longer-run outcomes. For example, skilled defense attorneys might be able to delay a case to be heard by a judge of the same-gender, which could erroneously lead to attributing the observed effects to the judge's gender. These results are presented for each of the three main outcomes in Online Appendix Table A10. Reassuringly, we do not find any impact of the juvenile defendant-judge gender interaction term, which is based on the gender of the presiding judge on the day of referral, on any of the outcomes.[29] In the second placebo test, we randomly assign gender to each judge in the sample. We repeat this exercise 1,000 times and display in Online Appendix Figure A4 the distribution of the coefficient estimates of the own-gender effect from this randomization inference procedure. In 18 out of 1,000 permutations, we find a coefficient estimate smaller than the actual one for incarceration (Panel A). Additionally, only five out of 1,000 permutations yield placebo estimates larger than the true estimates for education (Panel B), while 121 permutations show a smaller effect for TANF (Panel C).

6 Conclusion

A fundamental principle and ambition of the U.S. criminal justice system is one of fairness, i.e., in which all defendants are treated equally by the law. Motivated by fairness concerns, especially for populations that are under-represented amongst the agents in charge of executing the law, a large literature documents the existence of many types of judicial decision-making disparities. But is such unequal treatment only a concern from a fairness perspective? The answer to this question depends in part on whether in-court disparate treatment has significant out-of-court consequences and the size and direction of such consequences.

In the context of the South Carolina juvenile courts, we find that assignment to a same-gender judge as oneself significantly decreases the chance of incarceration. In fact, gender congruence in judicial assignment decreases the probability of incarceration by about 1.3 percentage points, which corresponds to a total of about 800 case decisions or more than 70 cases per year of our sample.[30] We also investigate the reasons behind these disparities in courtroom decisions. Our findings strongly suggest that judges engage in gender-based discrimination, with female judges being the primary drivers of this behavior. Furthermore, we provide the first evidence that such in-group judicial assignments have significant out-of-court consequences, including improvements in educational outcomes and adult welfare use and possibly lower overall criminal involvement. Using a similar back-of-the-envelope calculation, we find that gender congruence led to an

[29] A placebo test using the offense date to link case files to judges produces very similar results.

[30] These estimates are found by regressing the incarceration decision on an indicator for whether the judge and juvenile are the same gender, as well as our baseline controls (-0.013 with a s.e. = 0.006), and multiplying the number of case files (62,502) by -0.013. We must assume that the effect observed in our matched sample generalizes to the unmatched cases.

additional 700 juveniles enrolling in twelfth grade, while 575 fewer children received TANF as adults.

Our findings have important policy implications...(to be completed).

References

Agan, A. Y., J. L. Doleac, and A. Harvey (2023). Misdemeanor Prosecution. *Quarterly Journal of Economics 138*(3), 1453–1505.

Ahrsjö, U., S. Niknami, and M. Palme (2024). Identity in Court Decision-Making. *American Economic Journal: Economic Policy 16*(4), 142–164.

Aizer, A. and J. J. Doyle Jr. (2015). Juvenile Incarceration, Human Capital, and Future Crime: Evidence from Randomly Assigned Judges. *Quarterly Journal of Economics 130*(2), 759–803.

Akerlof, G. A. and R. E. Kranton (2000). Economics and Identity. *Quarterly Journal of Economics 115*(3), 715–753.

Alesina, A. and E. La Ferrara (2014). A Test of Racial Bias in Capital Sentencing. *American Economic Review 104*(11), 3397–3433.

Antonovics, K. and B. G. Knight (2009). A New Look at Racial Profiling: Evidence from the Boston Police Department. *Review of Economics and Statistics 91*(1), 163–177.

Anwar, S., P. Bayer, and R. Hjalmarsson (2012). The Impact of Jury Race in Criminal Trials. *Quarterly Journal of Economics 127*(2), 1017–1055.

Anwar, S., P. Bayer, and R. Hjalmarsson (2014). The Role of Age in Jury Selection and Trial Outcomes. *Journal of Law and Economics 57*(4), 1001–1030.

Anwar, S., P. Bayer, and R. Hjalmarsson (2019). Politics in the Courtroom: Political Ideology and Jury Decision Making. *Journal of the European Economic Association 17*(3), 834–875.

Anwar, S., P. Bayer, and R. Hjalmarsson (2022). Unequal Jury Representation and Its Consequences. *American Economic Review: Insights 4*(2), 159–174.

Anwar, S. and H. Fang (2006). An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence. *American Economic Review 96*(1), 127–151.

Arnold, D., W. Dobbie, and P. Hull (2022). Measuring Racial Discrimination in Bail Decisions. *American Economic Review 112*(9), 2992–3038.

Ash, E., S. Asher, A. Bhowmick, S. Bhupatiraju, D. Chen, T. Devi, C. Goessmann, P. Novosad, and B. Siddiqi (2023). In-Group Bias in the Indian Judiciary: Evidence from 5 Million Criminal Cases. *Working Paper*.

Bagues, M., M. Sylos-Labini, and N. Zinovyeva (2017). Does the Gender Composition of Scientific Committees Matter? *American Economic Review 107*(4), 1207–1238.

Baron, E. J., J. J. Doyle Jr, N. Emanuel, P. Hull, and J. Ryan (2024, aug). Discrimination in Multiphase Systems: Evidence from Child Protection*. *Quarterly Journal of Economics 139*(3), 1611–1664.

Baron, J. E., B. Jacob, and J. Ryan (2023). Pretrial Juvenile Detention. *Journal of Public Economics 217*, 104798.

Becker, G. S. (1957). *The Economics of Discrimination*. University of Chicago Press.

Bhuller, M., G. B. Dahl, K. V. Loken, and M. Mogstad (2020). Incarceration, Recidivism, and Employment. *Journal of Political Economy 128*(4), 1269–1324.

Bindler, A. and R. Hjalmarsson (2018). How Punishment Severity Affects Jury Verdicts: Evidence from Two Natural Experiments. *American Economic Journal: Economic Policy 10*(4), 36–78.

Bindler, A. and R. Hjalmarsson (2019). Path Dependency in Jury Decision Making. *Journal of the European Economic Association 17*(6), 1971–2017.

Bindler, A. and R. Hjalmarsson (2020). The Persistence of the Criminal Justice Gender Gap: Evidence from 200 Years of Judicial Decisions. *Journal of Law and Economics 63*(2), 297–339.

Boyd, C. L., L. Epstein, and A. D. Martin (2010). Untangling the Causal Effects of Sex on Judging. *American Journal of Political Science 54*(2), 389–411.

Cabral, M. and M. Dillender (2024). Gender Differences in Medical Evaluations: Evidence from Randomly Assigned Doctors. *American Economic Review 114*(2), 462–499.

Carrell, S. E., M. E. Page, and J. E. West (2010). Sex and Science: How Professor Gender Perpetuates the Gender Gap. *Quarterly Journal of Economics 125*(3), 1101–1144.

Chen, D. L., T. J. Moskowitz, and K. Shue (2016). Decision Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires. *Quarterly Journal of Economics 131*(3), 1181–1242.

Children's Law Center (2019). The Family Court Process for Children Charged with Criminal and Status Offenses: A Brief Overview of South Carolina's Juvenle Delinquency Proceedings.

Choi, D. D., A. Harris, and F. Shen-Bayh (2022). Ethnic Bias in Judicial Decision Making: Evidence from Criminal Appeals in Kenya. *American Political Science Review 116*(3), 1067–1080.

Cohen, A. and C. S. Yang (2019). Judicial Politics and Sentencing Decisions. *American Economic Journal: Economic Policy 11*(1), 160–191.

Dee, T. S. (2005). A Teacher Like Me: Does Race, Ethnicity, or Gender Matter? *American Economic Review 95*(2), 158–165.

Depew, B., O. Eren, and N. Mocan (2017). Judges, Juveniles, and In-Group Bias. *Journal of Law and Economics 62*(2), 209–239.

Eren, O. and N. Mocan (2018). Emotional Judges and Unlucky Juveniles. *American Economic Journal: Applied Economics 10*(3), 171–205.

Eren, O. and N. Mocan (2021). Juvenile Punishment, High School Graduation and Adult Crime: Evidence from Idiosyncratic Judge Harshness. *Review of Economics and Statistics 103*, 34–47.

Fairlie, R. W., F. Hoffmann, and P. Oreopoulos (2014). A Community College Instructor Like Me: Race and Ethnicity Interactions in the Classroom. *American Economic Review 104*(8), 2567–2591.

Fisman, R., D. Paravisini, and V. Vig (2017). Cultural Proximity and Loan Outcomes. *American Economic Review 107*(2), 457–492.

Flanagan, F. X. (2018). Race, Gender, and Juries: Evidence from North Carolina. *Journal of Law and Economics 61*(2), 189–214.

Garin, A., D. Koustas, C. McPherson, S. Norris, M. Pecenco, E. K. Rose, Y. Shem-Tov, and J. Weaver (2024). The Impact of Incarceration on Employment, Earnings, and Tax Filing. *NBER Working Paper 32747*.

Gazal-Ayal, O. and R. Sulitzeanu-Kenan (2010). Let My People Go: Ethnic In-Group Bias in Judicial Decisions: Evidence from a Randomized Natural Experiment. *Journal of Empirical Legal Studies 7*(3), 403–428.

Gershenson, S., C. M. D. Hart, J. Hyman, C. A. Lindsay, and N. W. Papageorge (2022). The Long-Run Impacts of Same-Race Teachers. *American Economic Journal: Economic Policy 14*(4), 300–342.

Goncalves, F. and S. Mello (2021). A Few Bad Apples? Racial Bias in Policing. *American Economic Review 111*(5), 1406–1441.

Guo, F., J. Li, M. Ma, and D. Zha (2023). Identity on the Bench: Gender In-Group Bias in the Judiciary. *Working Paper*.

Hester, R. (2017). Judicial Rotation as Centripetal Force: Sentencing in the Court Communities of South Carolina. *Criminology 55*(1), 205–235.

Hjalmarsson, R. and L. Lochner (2012). The Impact of Education on Crime: International Evidence. *CESifo DICE Report 10*(2), 49–55.

Hoekstra, M. and C. Sloan (2022). Does Race Matter for Police Use of Force? Evidence from 911 Calls. *American Economic Review 112*(3), 827–860.

Hoekstra, M. and B. Street (2021). The Effect of Own-Gender Jurors on Conviction Rates. *Journal of Law and Economics 64*(3), 513–537.

Hoffmann, F. and P. Oreopoulos (2009). A Professor Like Me: The Influence of Instructor Gender on College Achievement. *Journal of Human Resources 44*(2), 479–494.

Kling, J. R. (2006). Incarceration Length, Employment, and Earnings. *American Economic Review 96*(3), 863–876.

Lim, C. S. H., B. S. Silveira, and J. M. Snyder (2016). Do Judges Characteristics Matter? Ethnicity, Gender, and Partisanship in Texas State Trial Courts. *American Law and Economics Review 18*(2), 302–357.

Lim, C. S. H., J. M. Snyder Jr., and D. Strömberg (2015). The Judge, the Politician, and the Press: Newspaper Coverage and Criminal Sentencing across Electoral Systems. *American Economic Journal: Applied Economics 7*(4), 103–135.

Linton, O., E. Maasoumi, and Y.-J. Whang (2005). Consistent Testing for Stochastic Dominance under General Sampling Schemes. *Review of Economic Studies 72*(3), 735–765.

Loeffler, C. E. and D. S. Nagin (2022). The Impact of Incarceration on Recidivism. *Annual Review of Criminology 5*(1), 133–152.

Mueller-Smith, M. (2015). The Criminal and Labor Market Impacts of Incarceration. *Working Paper*.

National Center for State Courts (2014). State of the State Courts.

National Center for State Courts (2023). State of the State Courts.

Norris, S., M. Pecenco, and J. Weaver (2021). The Effects of Parental and Sibling Incarceration: Evidence from Ohio. *American Economic Review 111*(9), 2926–2963.

Oreopoulos, P. and K. G. Salvanes (2011). Priceless: The Nonpecuniary Benefits of Schooling. *Journal of Economic Perspectives 25*(1), 159–184.

Parsons, C. A., J. Sulaeman, M. C. Yates, and D. S. Hamermesh (2011). Strike Three: Discrimination, Incentives, and Evaluation. *American Economic Review 101*(4), 1410–1435.

Philippe, A. and A. Ouss (2018). No Hatred or Malice, Fear or Affection: Media and Sentencing. *Journal of Political Economy 126*(5), 2134–2178.

Price, J. and J. Wolfers (2010). Racial Discrimination Among NBA Referees. *Quarterly Journal of Economics 125*(4), 1859–1887.

Puzzanchera, C., S. Hockenberry, and M. Sickmund (2022). Youth and the Juvenile Justice System: 2022 National Report. *National Center for Juvenile Justice*.

Smith, M., Hyman. Michael B, and S. E. Redfield (2014). Addressing Bias Among Judges. *State Court Report*.

South Carolina Department of Juvenile Justice (2017). Data Resource Guide.

Ye, H. and J. Yi (2023). Patient-Physician Race Concordance, Physician Decisions, and Patient Outcomes. *Review of Economics and Statistics 105*(4), 766–779.

Zeltzer, D. (2020). Gender Homophily in Referral Networks: Consequences for the Medicare Physician Earnings Gap. *American Economic Journal: Applied Economics 12*(2), 169–197.

Table 1: Summary Statistics for Juveniles and Judges

	Matched Sample		Unmatched Sample	
	Mean	SD	Mean	SD
	(1)	(2)	(3)	(4)
Panel A: Juvenile Characteristics				
Incarceration	0.142	0.349	0.156	0.363
Probation	0.738	0.439	0.755	0.430
Acquitted	0.119	0.324	0.088	0.284
Female	0.281	0.449	0.288	0.453
Black	0.561	0.496	0.577	0.494
White	0.409	0.492	0.382	0.486
Age at Disposition	15.31	1.62	15.48	1.45
Most Serious Charge:				
Property Offense	0.209	0.407	0.188	0.391
Violent Offense	0.156	0.362	0.159	0.367
School-Related Offense	0.221	0.415	0.195	0.396
Charge Severity	3.395	3.490	3.350	3.341
Charged with Multiple Offenses	0.324	0.468	0.333	0.472
Adult Outcomes:				
Enrolled in 12th Grade	0.168	0.374	0.191	0.393
Graduated High School within 4 Years	0.140	0.347	0.151	0.359
TANF Receipt	0.137	0.344	0.131	0.337
SNAP Receipt	0.798	0.402	0.755	0.430
Adult Arrest	0.630	0.483	0.635	0.481
Sample Size	23,569		38,933	
Panel B: Judge Characteristics				
Female	0.264	0.441		
Age at Disposition	53.32	8.70		
White	0.948	0.222		
Graduated from a College in the State	0.855	0.352		
Judge Experience in Legal Practice	22.80	8.00		
Number of Judges	84			
Panel C: County Characteristics				
Number of Family Court Judges	6.98	3.92	13.92	5.83
Log of County Population	11.10	0.74	12.41	0.60
Unemployment Rate (%)	9.62	3.49	7.43	2.41

NOTES: This table reports summary statistics for the samples of juvenile case files that can and cannot be linked to disposition judges over the period from 2002 to 2012. Juvenile case files can be successfully matched to judges using the disposition date when cases are handled by a single family court judge on any given day. The statistics for adult outcomes in Panel A are weighted by the inverse of the number of times a juvenile defendant is observed in the data.

Table 2: Randomization and Prediction Tests

	Incarceration	Acquittal	Female Judge	Minority Judge	Judge Age	Judge Leniency
	Coefficients (Standard Errors)					
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.046***	-0.009*	0.002	0.000	-0.010	-0.000
	(0.007)	(0.005)	(0.003)	(0.002)	(0.046)	(0.000)
Black	0.034***	0.008**	0.001	-0.001	-0.035	-0.000
	(0.008)	(0.004)	(0.003)	(0.001)	(0.061)	(0.002)
White	-0.034***	-0.007	-0.001	0.000	0.088	0.000
	(0.008)	(0.004)	(0.003)	(0.001)	(0.054)	(0.000)
Age at Disposition	0.026***	0.003	-0.000	0.000	0.049***	0.000
	(0.003)	(0.002)	(0.001)	(0.001)	(0.019)	(0.000)
Property Offense	-0.021*	0.017**	0.006	0.002	-0.129	0.000
	(0.011)	(0.008)	(0.004)	(0.002)	(0.081)	(0.000)
Violent Offense	-0.023***	0.075***	-0.001	0.001	0.057	0.000
	(0.007)	(0.017)	(0.005)	(0.002)	(0.058)	(0.000)
School-Related Offense	-0.136***	-0.048***	-0.004	-0.002	-0.037	0.000
	(0.014)	(0.009)	(0.005)	(0.002)	(0.083)	(0.000)
Offense Severity	0.009***	0.007***	-0.000	0.000	0.001	0.000
	(0.002)	(0.002)	(0.001)	(0.001)	(0.009)	(0.000)
Number of Charges at Disposition	0.026***	-0.009***	0.000	-0.000	-0.001	0.000
	(0.004)	(0.003)	(0.001)	(0.001)	(0.024)	(0.000)
Joint Significance (p-value)	0.00	0.00	0.54	0.93	0.16	0.73
Sample Size	23,569	23,569	23,569	18,090	23,108	23,569

NOTES: Each cell represents a separate regression and all regression estimations in Columns (1)-(5) control for court-by-year-month and court-by-day-of-week fixed effects. The dependent variable in the last column is the leave-one-out average non-incarceration rate. We construct the leave-one-out measure using the residuals from a regression of the non-incarceration indicator on court-by-time fixed effects. To allow for meaningful variation in judge leniency measure, court-by-time fixed effects in Column (6) include court-by-year-by-quarter fixed effects. Standard errors are two-way clustered at the judge and court levels.
* significant at 10%, ** significant at 5%, *** significant at 1%.

Table 3: The Effect of Juvenile Defendant-Judge Gender Interactions on Judicial Dispositions

	Incarceration			Acquitted	Predicted Incar.
	Coefficients (Standard Errors)				
	(1)	(2)	(3)	(4)	(5)
Female Judge	0.020	0.021			
	(0.015)	(0.013)			
Female Defendant*Female Judge	-0.024**	-0.026**	-0.026**	0.001	0.002
	(0.012)	(0.012)	(0.012)	(0.009)	(0.004)
Outcome Mean	0.142			0.119	
Sample Size	23,569	23,569	23,569	23,569	23,569
Controls:					
Court-by-Time Fixed Effects	Yes	Yes	Yes	Yes	Yes
Juvenile Characteristics	No	Yes	Yes	Yes	No
Judge Fixed Effects	No	No	Yes	Yes	Yes

NOTES: All regression estimations control for court-by-year-month and court-by-day-of-week fixed effects. Juvenile controls include indicators for gender and race, birth year fixed effects, severity of the offense, offense fixed effects, number of charges at disposition, and number of times a juvenile defendant is observed in the data. The outcome variable in Column (5) is obtained from a regression of incarceration indicator on observable juvenile characteristics. Standard errors are two-way clustered at the judge and court levels. ** significant at 5%, *** significant at 1%.

Table 4: Estimates of Mean Recidivism Risk and Disparate Impact of Non-Incarceration Decisions

	Local Linear Extrapolations	Linear Extrapolations
	(1)	(2)
Panel A: Juvenile Recidivism Risk		
Female Defendants	0.109***	0.102***
	(0.010)	(0.007)
Male Defendants	0.139***	0.117***
	(0.013)	(0.008)
Panel B: Disparate Impact		
Average Across Judges	-0.012**	-0.014***
	(0.006)	(0.005)

NOTES: This table reports estimates of mean recidivism risk and disparate impact of non-incarceration decisions. Column (1) reports the estimates from local linear extrapolations using a Gaussian kernel with gender-specific rule-of-thumb bandwidth, while Column (2) presents the results from linear extrapolations. Robust standard errors are two-way clustered at the judge and court levels and are obtained by a bootstrapping procedure, as described in Arnold et al. (2022).
** significant at 5%, *** significant at 1%.

Table 5: Regressions of Judge-Level Disparity Impacts on Judge Characteristics

	Coefficients (Standard Errors)			
	(1)	(2)	(3)	(4)
Female Judge	-0.018	-0.020	-0.020	-0.017
	(0.013)	(0.014)	(0.014)	(0.016)
Judges with Above Average Experience		-0.008	-0.008	-0.004
		(0.012)	(0.012)	(0.015)
Residential Jurisdiction-Urban County			0.000	-0.004
			(0.014)	(0.016)
White Judge				0.026
				(0.031)
Number of Judges	83	81	81	67
Control Mean-Disparate Impact:				
Male Judges	-0.007			
Judges with Below Average Experience	-0.012			
Residential Jurisdiction-Rural County	-0.011			
Minority Judges	-0.049			

NOTES: This table reports the OLS estimates of judge-level disparity impacts (based on local linear extrapolations) on judge characteristics. Judges with above average experience in legal practice is defined as judges with legal experience above the sample average (22 years). Urban county is an indicator for whether the main jurisdiction of the judge is a county with above average population size. Robust standard errors are two-way clustered at the judge and court levels and are obtained by a bootstrapping procedure, as described in Arnold et al. (2022).

Table 6: The Effect of Juvenile Defendant-Judge Gender Interactions on Educational and Welfare Outcomes

	Took the HS Exit Exams	Enrolled in 12th Grade	Graduate from HS in 4 Years	Schooling Index	TANF Receipt	SNAP Receipt
	Coefficients (Standard Errors)					
	(1)	(2)	(3)	(4)	(5)	(6)
Female Defendant*Female Judge	0.042***	0.039**	0.022*	0.105***	-0.032	-0.001
	(0.016)	(0.016)	(0.013)	(0.031)	(0.026)	(0.016)
Outcome Mean	0.412	0.168	0.140	0.000	0.137	0.798
Sample Size	15,416	23,569	17,629	23,569	23,569	23,569
Controls:						
Court-by-Time Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Juvenile Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Judge Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes

NOTES: All regression estimations control for court-by-year-month and court-by-day-of-week fixed effects. Juvenile controls include indicators for gender and race, birth year fixed effects, severity of the offense, offense fixed effects, and number of charges at disposition. Column (1) and (4) limit the analysis sample to birth cohorts (1990-1998) that were impacted by the state's exit exam policy and to those juveniles who took the tests after the disposition date. The schooling index is constructed by averaging z-scores over all non-missing components. Standard errors are two-way clustered at the judge and court levels. All regressions are weighted by the inverse of the number of times a juvenile defendant is observed in the data.

*significant at 10%, ** significant at 5%, *** significant at 1%.

Table 7: The Effect of Juvenile Defendant-Judge Gender Interactions on Judicial Dispositions and Longer-Run Outcomes-By Judge's Experience in Legal Practice

	Incarceration	Acquitted	Schooling Index	TANF Receipt	Adult Crime
			Coefficients (Standard Errors)		
	(1)	(2)	(3)	(4)	(5)
Panel A: Judge's Exper. in Legal Practice>=22 Years					
Female Defendant*Female Judge	-0.047***	0.026	0.268***	-0.068**	-0.016
	(0.011)	(0.019)	(0.061)	(0.031)	(0.027)
Outcome Mean	0.150	0.108	-0.016	0.137	0.657
Sample Size	14,000	14,000	14,000	14,000	14,000
Panel B: Judge's Exper. in Legal Practice<22 Years					
Female Defendant*Female Judge	-0.020	0.006	0.057	0.006	-0.002
	(0.021)	(0.012)	(0.052)	(0.036)	(0.042)
Outcome Mean	0.131	0.139	-0.020	0.138	0.614
Sample Size	9,318	9,318	9,318	9,318	9,318

NOTES: A judge's professional experience in law is defined as the difference between year the judge is first observed in the data and year of graduating from law school. The sample average is 22 years. All regression estimations control for court-by-year-month and court-by-day-of-week fixed effects. Juvenile controls include indicators for gender and race, birth year fixed effects, severity of the offense, offense fixed effects, and number of charges at disposition. Standard errors are two-way clustered at the judge and court levels. All regressions in Columns 3-5 are weighted by the inverse of the number of times a juvenile defendant is observed in the data.

** significant at 5%, *** significant at 1%.

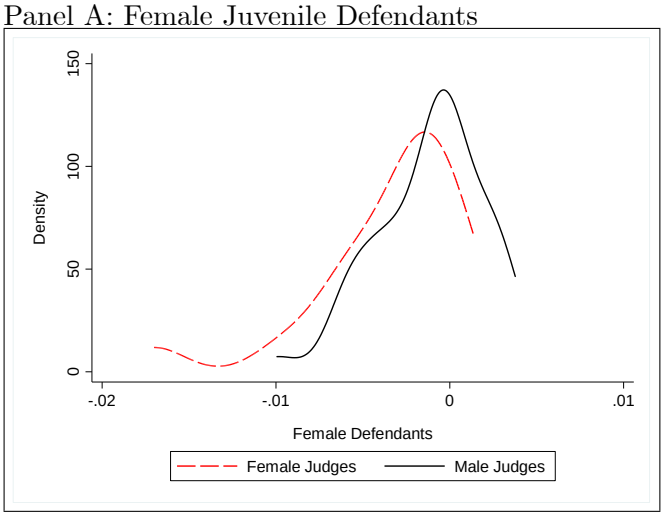
Table 8: Robustness Checks

	Add More Judge Def. Interaction Terms	Control Court-by-Year-Week Fixed Effects	Control Court-by-Year-Week-Day Fixed Effects	Add Detailed Offense Fixed Effects	Mandatory Sch Attendance Disp. Excluded
			Coefficients (Standard Errors)		
	(1)	(2)	(3)	(4)	(5)
Panel A: Incarceration					
Female Defendant*Female Judge	-0.030***	-0.022*	-0.027**	-0.023*	-0.040**
	(0.010)	(0.012)	(0.012)	(0.012)	(0.016)
Sample Size	17,878	23,569	23,569	23,569	19,679
Panel B: Schooling Index					
Female Defendant*Female Judge	0.103***	0.112***	0.123***	0.106***	0.098**
	(0.036)	(0.031)	(0.033)	(0.033)	(0.044)
Sample Size	17,878	23,569	23,569	23,569	19,679
Panel C: TANF Receipt					
Female Defendant*Female Judge	-0.048*	-0.037	-0.042	-0.033	-0.045
	(0.025)	(0.026)	(0.027)	(0.026)	(0.031)
Sample Size	17,878	23,569	23,569	23,569	19,679
Controls:					
Court-by-Time Fixed Effects	Yes	Yes	Yes	Yes	Yes
Juvenile Characteristics	Yes	Yes	Yes	Yes	Yes
Judge Fixed Effects	Yes	Yes	Yes	Yes	Yes

NOTES: The first column controls for interactions between female judge and juvenile characteristics (race, birth year fixed effects, severity of the offense, offense fixed effects and number of charges at disposition) and interactions between female defendant and judge characteristics (race, information on alma mater, and years of experience). Column (2) reports defendant-judge female interaction effect from a model which conditions on court-by-year-week and court-by-day-of-week fixed effects, while Column (3) presents the results after controlling for court-by-year-week-day fixed effects. The specifications in all other columns control for court-by-year-month and court-by-day-of-week fixed effects. Column (4) adds detailed offense fixed effects to the baseline specification. There are 350 uniquely defined juvenile offense types in the analysis sample. The last column excludes all non-incarceration convictions involving mandatory school attendance. Standard errors are two-way clustered at the judge and court levels. Juvenile controls include indicators for gender and race, birth fixed effects, severity of the offense, offense fixed effects and number of charges at disposition. Panel A additionally controls for the number of times a juvenile defendant is observed in the data. The regressions in Panels B and C are weighted by the inverse of the of times a juvenile defendant is observed in the data.

* significant at 10%, ** significant at 5%, *** significant at 1%.

Panel A: Female Juvenile Defendants



Panel B: Male Juvenile Defendants

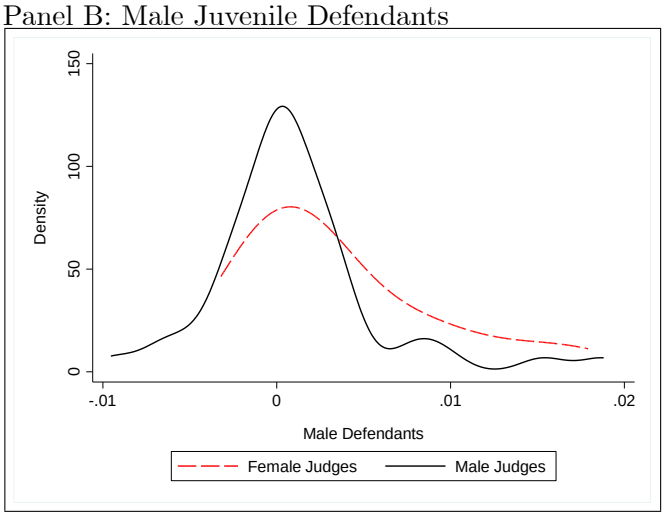
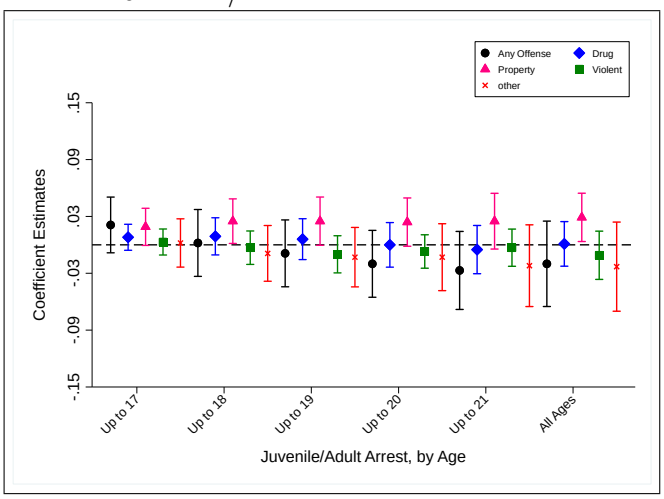


Figure 1: Individual Judge Effects-Empirical Bayes Estimates

NOTES: Distribution of the individual judge effects by juvenile defendant's gender. We first regress the incarceration indicator on court-by-time fixed effects, along with case, juvenile, and observable judge characteristics, and retain the residuals. We then use these residuals to estimate a random effect model, obtaining empirical Bayes estimates for each individual judge by defendant gender. For ease of exposition, the x-axis is multiplied by 10^{15}.

Panel A: Juvenile/Adult Arrest



Panel B: Adult Arrest

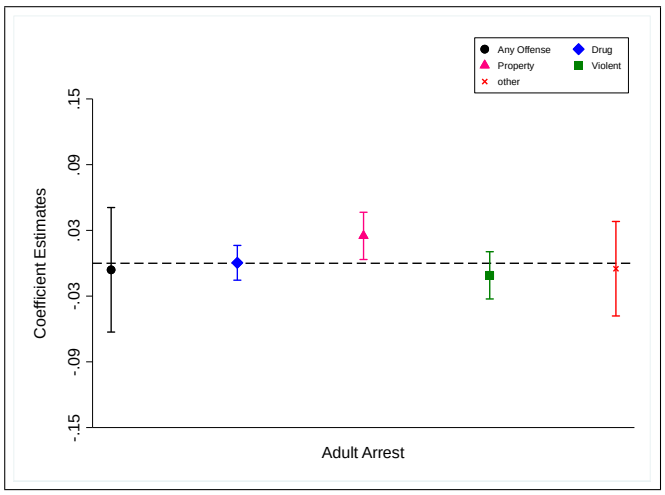


Figure 2: Effect of Juvenile Defendant-Judge Gender Interactions on the Probability of Ever Being Arrested

NOTES: The figure plots the effect of juvenile defendant-judge gender interactions on the probability of ever being arrested. Each point in Panel A comes from a separate regression, using samples that increase age moving rightward along the x-axis. The analysis sample in Panel A is restricted to first-time dispositions and the dependent variable is an indicator equal to one if the juvenile was first arrested of an offense by the given age. The analysis sample in Panel B includes all dispositions and the dependent variable is an indicator equal to one if the juvenile was arrested of an offense as an adult. Confidence intervals (95%) are based on standard errors that are two-way clustered at the judge and court levels.

Online Appendix A:

Table A1: Summary Statistics by Juvenile's Gender

	Female Juveniles		Male Juveniles	
	Mean	SD	Mean	SD
	(1)	(2)	(3)	(4)
Juvenile/County Characteristics				
Incarceration	0.100	0.300	0.158	0.365
Probation	0.797	0.402	0.715	0.451
Acquitted	0.102	0.303	0.126	0.332
Black	0.487	0.500	0.590	0.491
White	0.484	0.500	0.380	0.485
Age at Disposition	15.20	1.71	15.36	1.58
Most Serious Charge:				
Property Offense	0.102	0.303	0.251	0.434
Violent Offense	0.139	0.346	0.162	0.368
School-Related Offense	0.311	0.463	0.186	0.389
Charge Severity	2.229	2.324	3.852	3.754
Charged with Multiple Offenses	0.236	0.424	0.359	0.480
Adult Outcomes:				
Enrolled in 12th Grade	0.207	0.405	0.151	0.358
Graduated High School within 4 Years	0.187	0.390	0.121	0.326
TANF Receipt	0.329	0.470	0.056	0.230
SNAP Receipt	0.873	0.334	0.766	0.423
Adult Arrest	0.489	0.500	0.689	0.463
Sample Size	6,630		16,939	

NOTES: This table reports summary statistics for the samples of juvenile case files that can be linked to disposition judges over the period from 2002 to 2012. Juvenile case files can be successfully matched to judges using the disposition date when cases are handled by a single family court judge on any given day. The statistics for adult outcomes are weighted by the inverse of the number times a juvenile defendant is observed in the data.

Table A2: Missingness of Judge and Defendant/Case Characteristics

	Full Sample	Full Sample	Females	Males
	Coefficients (Standard Errors)			
	(1)	(2)	(3)	(4)
Incarceration		-0.0026	0.0044	-0.0047
		(0.0054)	(0.0052)	(0.0061)
Female	0.0021	0.0020		
	(0.0029)	(0.0029)		
Black	0.0005	0.0005	-0.0031	0.0014
	(0.0021)	(0.0021)	(0.0027)	(0.0026)
Age at Disposition	-0.0002	-0.0001	0.0000	-0.0002
	(0.0007)	(0.0007)	(0.0018)	(0.0009)
Property Offense	0.0018	0.0015	0.0037	0.0010
	(0.0033)	(0.0031)	(0.0074)	(0.0035)
Violent Offense	0.0015	0.0012	0.0036	0.0012
	(0.0034)	(0.0032)	(0.0093)	(0.0040)
School-Related Offense	0.0116**	0.0112**	0.0167**	0.0108**
	0.0053	(0.0052)	(0.0077)	(0.0052)
Offense Severity	-0.0005	-0.0004	-0.0001	-0.0004
	(0.0004)	(0.0004)	(0.0011)	(0.0004)
Number of Charges at Disposition	-0.0007	-0.0006	0.0013	-0.0014
	(0.0010)	(0.0010)	(0.0022)	(0.0011)
Joint Significance (p-value)	0.34	0.42	0.57	0.41
Sample Size	62,502	62,502	17,860	44,642

NOTES: All regression estimations control for court-by-year-month and court-by-day-of-week fixed effects. The dependent variable takes the value one if the juvenile defendant is not successfully matched to the disposition judge. Standard errors are clustered at the court level. ** significant at 5%.

Table A3: Estimates of Disparate Impact of Non-Incarceration Decisions: Alternative Timeframes of Juvenile Recidivism

	Disparate Impact
Panel A: Three-Months Juvenile Recidivism	
Average Across Judges	-0.017***
	(0.006)
Panel B: One-Year Juvenile Recidivism	
Average Across Judges	-0.013**
	(0.006)

NOTES: This table reports estimates of disparate impact of non-incarceration decisions using alternative timeframes of juvenile recidivism. Statistical extrapolations are based on local linear extrapolations with a Gaussian kernel and gender-specific rule-of-thumb bandwidths. Robust standard errors are two-way clustered at the judge and court levels and are obtained by a bootstrapping procedure, as described in Arnold et al. (2022). ** significant at 5%, *** significant at 1%.

Table A4: Summary Statistics-Juvenile and Adult Arrest

	Mean	SD
	(1)	(2)
Panel A: Adult Arrest		
Adult Arrest-Any Offense	0.630	0.483
Adult Arrest-Drug Offense	0.129	0.336
Adult Arrest-Property Offense	0.164	0.370
Adult Arrest-Violent Offense	0.089	0.285
Adult Arrest-Other Offense	0.222	0.416
Sample Size	23,569	
Panel B: Juvenile/Adult Arrest (After First Disposition)		
Juvenile/Adult Arrest-Any Offense	0.694	0.461
Juvenile/Adult Arrest-Drug Offense	0.139	0.346
Juvenile/Adult Arrest-Property Offense	0.197	0.398
Juvenile/Adult Arrest-Violent Offense	0.110	0.313
Juvenile/Adult Arrest-Other Offense	0.344	0.475
Sample Size	15,951	

NOTES: This table reports summary statistics for the sample of juvenile case files that can be linked to disposition judges over the period from 2002 to 2012. Juvenile case files can be successfully matched to judges using the disposition date when cases are handled by a single family court judge on any given day. The statistics in Panel A are weighted by the inverse number of times a juvenile defendant is observed in the data. The statistics in Panel B includes juvenile and adult arrests and the analysis sample is restricted to first-time dispositions.

Table A5: Summary Statistics-Judge Characteristics by Judge's Experience in Legal Practice

	Judge Experience>=22 years		**Judge Experience<22 Years**	
	Mean	SD	Mean	SD
	(1)	(2)	(3)	(4)
Female	0.139	0.346	0.456	0.498
Age at First Time Observed	54.54	4.51	41.16	6.62
Age at Disposition	58.63	4.87	45.45	7.06
White	0.964	0.185	0.920	0.272
Graduated from a College in the State	0.846	0.361	0.865	0.342
Number of Juvenile Case Files	14,000		9,318	
Number of Judges	56		26	

NOTES: A judge's professional experience in law is defined as the difference between year the judge is first observed in the data and year of graduating from law school. The sample average is 22 years.

Table A6: The Effect of Juvenile Defendant-Judge Gender Interactions on Judicial Dispositions-Different Time Windows

	Number of Days Since Referral				Days Since Prosecution	
	Days<21	**Days>=21**	**Days<10**	**Days>=10**	**Days<7**	**Days>=7**
	Coefficients (Standard Errors)					
	(1)	(2)	(3)	(4)	(5)	(6)
Female Defendant*Female Judge	-0.043*	-0.020*	-0.012	-0.030**	-0.020	-0.029**
	(0.026)	(0.011)	(0.050)	(0.011)	(0.038)	(0.012)
Outcome Mean	0.247	0.118	0.282	0.125	0.226	0.126
Sample Size	4,437	19,132	2,616	20,953	3,692	19,877
Controls:						
Court-by-Time Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Juvenile Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Judge Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes

NOTES: Columns (1)-(4) partition the analysis sample based on the number of days since initial referral to the juvenile justice system. Columns (5) and (6) divide the sample based on the number of elapsed days relative to prosecution. Juvenile controls include indicators for gender and race, birth year fixed effects, severity of the offense, offense fixed effects, number of charges at disposition, and number of times a juvenile defendant is observed in the data. Standard errors are two-way clustered at the judge and court levels.

* significant at 10%, ** significant at 5%.

Table A7: Robustness Checks-by Different Weights

	One Time Disposition	No Weighting
	(3)	(4)
Panel A: Incarceration		
Female Defendant*Female Judge	-0.020*	-0.026**
	(0.010)	(0.012)
Outcome Mean	0.090	0.142
Sample Size	11,094	23,569
Panel B: Schooling Index		
Female Defendant*Female Judge	0.105***	0.091**
	(0.037)	(0.036)
Outcome Mean	0.071	0.000
Sample Size	11,094	23,569
Panel C: TANF Receipt		
Female Defendant*Female Judge	-0.033	-0.021
	(0.032)	(0.024)
Outcome Mean	0.133	0.141
Sample Size	11,094	23,569
Controls:		
Court-by-Time Fixed Effects	Yes	Yes
Juvenile Characteristics	Yes	Yes
Judge Fixed Effects	Yes	Yes

NOTES: All regression estimations control for court-by-year-month and court-by-day-of-week fixed effects. Juvenile controls include indicators for gender and race, birth year fixed effects, severity of the offense, offense fixed effects, number of charges at disposition and number of times a juvenile defendant is observed in the data. Standard errors are two-way clustered at the judge and court levels.
*significant at 10%, ** significant at 5%, *** significant at 1%.

Table A8: The Effect of Juvenile Defendant-Judge Gender Interactions on Judicial Dispositions and Longer-Run Outcomes-Leave-One Juvenile Crime Type Out

	Leave-One Juvenile Crime Type Out				
	Drug Offense Excl.	**Property Offense Excl.**	**Violent Offense Excl.**	**Other Offense Excl.**	**School-Related Offense Excl.**
	Coefficients (Standard Errors)				
	(1)	(2)	(3)	(4)	(5)
Panel A: Incarceration					
Female Defendant*Female Judge	-0.023*	-0.025**	-0.024**	-0.013*	-0.036**
	(0.012)	(0.012)	(0.012)	(0.008)	(0.018)
Sample Size	22,095	18,633	19,901	15,289	18,358
Panel B: Schooling Index					
Female Defendant*Female Judge	0.113***	0.116***	0.102***	0.096**	0.077*
	(0.035)	(0.033)	(0.032)	(0.042)	(0.043)
Sample Size	22,095	18,633	19,901	15,289	18,358
Panel C: TANF Receipt					
Female Defendant*Female Judge	-0.027	-0.022	-0.037	-0.034	-0.029
	(0.026)	(0.028)	(0.027)	(0.025)	(0.031)
Sample Size	22,095	18,633	19,901	15,289	18,358
Controls:					
Court-by-Time Fixed Effects	Yes	Yes	Yes	Yes	Yes
Juvenile Characteristics	Yes	Yes	Yes	Yes	Yes
Judge Fixed Effects	Yes	Yes	Yes	Yes	Yes

NOTES: All regression estimations control for court-by-year-month and court-by-day-of-week fixed effects. Juvenile controls include indicators for gender and race, birth year fixed effects, severity of the offense, offense fixed effects, and number of charges at disposition. Panel A additionally controls for the number of times a juvenile defendant is observed in the data. The regressions in Panels B and C are weighted by the inverse of the times juvenile defendant is observed in the data. Standard errors are tow-way clustered at the judge and court levels.
*significant at 10%, ** significant at 5%, *** significant at 1%.

Table A9: Robustness to Alternative Levels of Clustering

	Clustered at the Court Level	Clustered at the Child Level	Clustered at the Unit of Randomization
	Coefficients (Standard Errors)		
	(1)	(2)	(3)
Panel A: Incarceration			
Female Defendant*Female Judge	-0.026**	-0.026**	-0.026**
	(0.010)	(0.011)	(0.011)
Panel B: Schooling Index			
Female Defendant*Female Judge	0.105**	0.105**	0.105**
	(0.041)	(0.043)	(0.043)
Panel C: TANF Receipt			
Female Defendant*Female Judge	-0.032	-0.032**	-0.032*
	(0.025)	(0.016)	(0.017)
Controls:			
Court-by-Time Fixed Effects	Yes	Yes	Yes
Juvenile Characteristics	Yes	Yes	Yes
Judge Fixed Effects	Yes	Yes	Yes

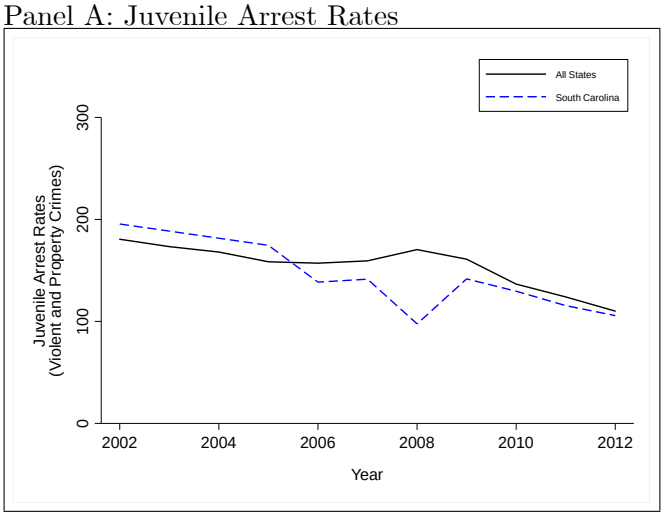
NOTES: All regression estimations control for court-by-year-month and court-by-day-of-week fixed effects. Juvenile controls include indicators for gender and race, birth year fixed effects, severity of the offense, offense fixed effects and number of charges at disposition. Panel A additionally controls for the number of times a juvenile defendant is observed in the data. All regressions in Panels B and C are weighted by the inverse of the number of times a juvenile defendant is observed in the data. *significant at 10%, ** significant at 5%.

Table A10: Placebo Effect of Juvenile Defendant-Judge Gender Interactions on Judicial Dispositions and Longer-Run Outcomes

	Incarceration	Schooling Index	TANF Receipt
	Coefficients (Standard Errors)		
	(1)	(2)	(5)
Female Defendant*Female Judge	0.003	0.010	-0.003
	(0.011)	(0.043)	(0.023)
Outcome Mean	0.149	0.000	0.135
Sample Size	18,353	18,353	18,353
Controls:			
Court-by-Time Fixed Effects	Yes	Yes	Yes
Juvenile Characteristics	Yes	Yes	Yes
Judge Fixed Effects	Yes	Yes	Yes

NOTES: Juveniles case files are matched to judges using the referral date. All regression estimations control for court-by-year-month and court-by-day-of-week fixed effects. Juvenile controls include indicators for gender and race, birth year fixed effects, severity of the offense, offense fixed effects, and number of charges at disposition. Standard errors are two-way clustered at the judge and court levels. The regressions in Columns (2) and (3) are weighted by the inverse of the number of times a juvenile defendant is observed in the data. All regressions in Columns 2 and 3

Panel A: Juvenile Arrest Rates



Panel B: Prosecution Rates

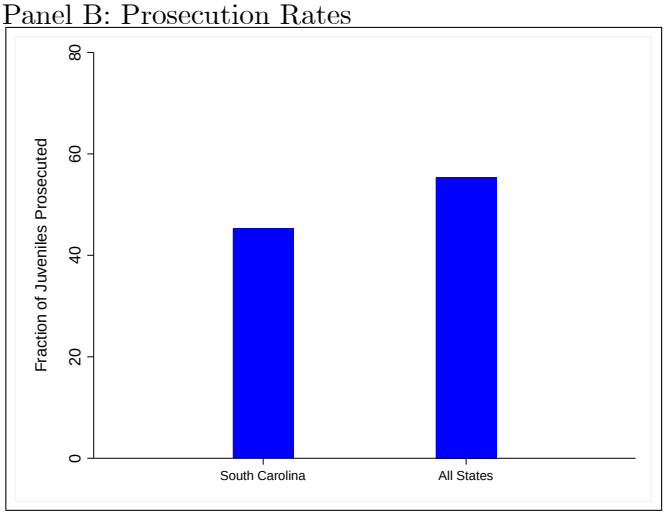
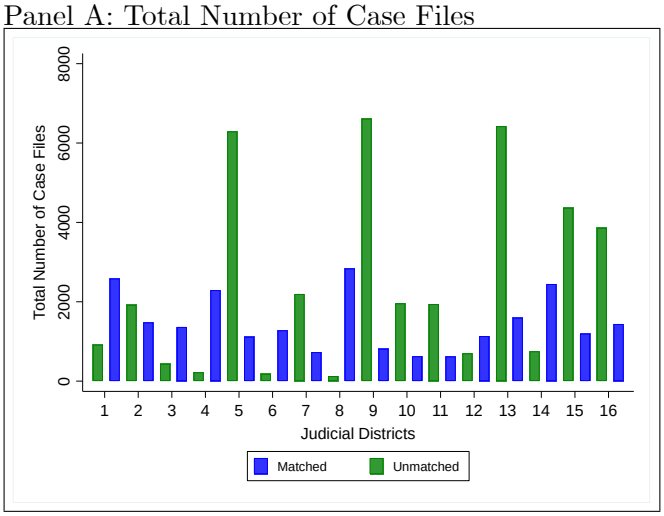


Figure A1: Youth Arrest Rates and Fraction of Juveniles Prosecuted

NOTES: Panel A plots juvenile arrest rates for violent and property crimes combined from 2002-2012, defined as the number of arrests of individuals under age 18 for every 10,000 persons aged 10-17. Panel B presents the fraction of juvenile referrals that resulted in prosecution over the same period. The youth crime statistics (Panel A) and national estimates on the fraction prosecuted (Panel B) are drawn from the periodical Juvenile Court Statistics reports.

Panel A: Total Number of Case Files



Panel B: Fraction of Matched Case Files

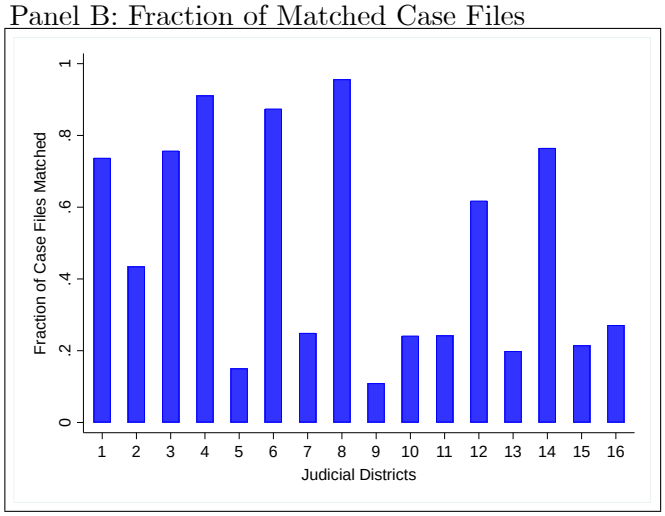


Figure A2: Matched and Unmatched Case Files by Judicial Districts

NOTES: The figure plots the total number (Panel A) and fraction of case files (Panel B) by judicial districts that can be linked to disposition judges over the period from 2002 to 2012. Juvenile case files can be successfully matched to judges using the disposition date when cases are handled by a single family court judge on any given date. The judicial system in structured into 16 judicial districts, each comprising a combination of state's 46 counties.

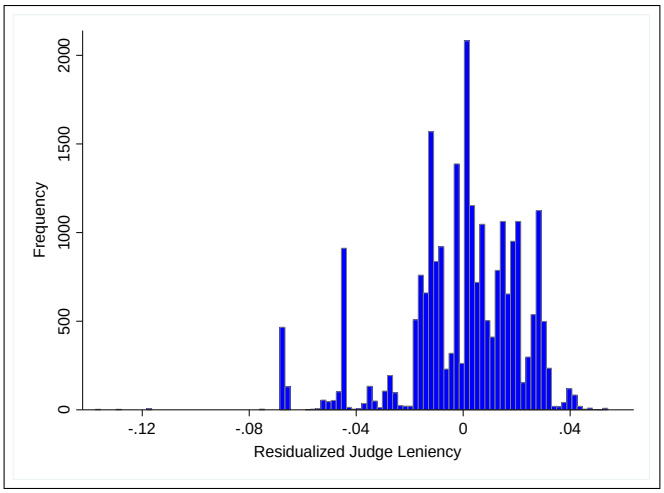
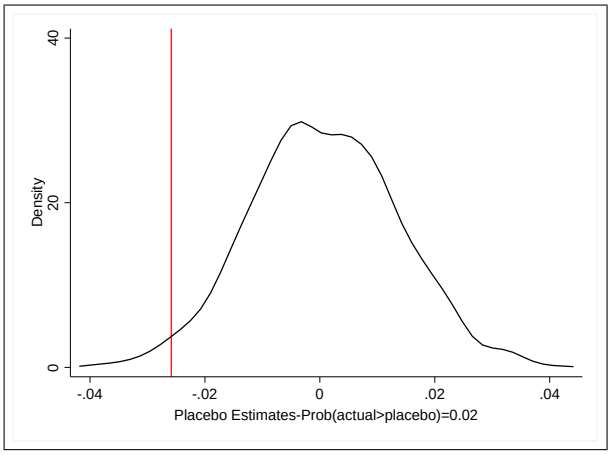


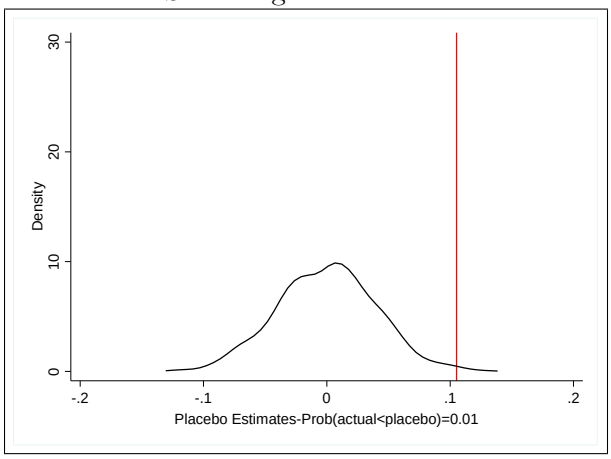
Figure A3: Distribution of Judge Leniency

NOTES: The histogram displays the distribution of the residualized judge leniency, obtained by the leave-one-out average non-incarceration rate among all other juvenile defendants assigned to a judge. We construct the leave-one-out measure using the residuals from a regression of the non-incarceration indicator on court-by-time fixed effects.

Panel A: Incarceration



Panel B: Schooling Index



Panel C: TANF Receipt

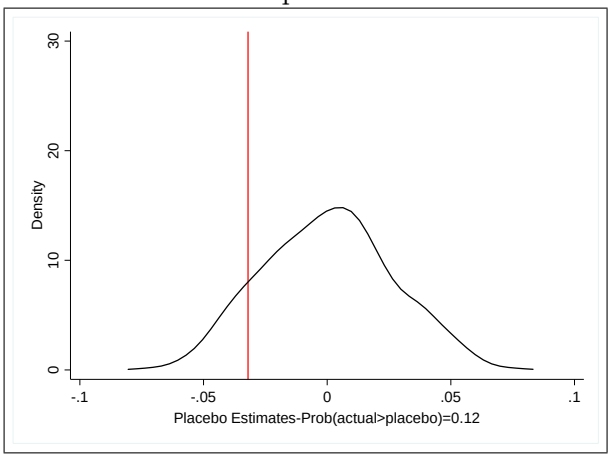


Figure A4: Placebo Estimates of the Effect of Juvenile Defendant-Judge Gender Interactions on the Incarceration Decision of Judges and Short- and Long-Run Outcomes

NOTES: Distribution of the coefficient estimates resulting from 1,000 sets of random assignment of gender to each judge in the analysis sample. The vertical lines denote the actual estimates. The fraction of placebo estimates smaller (larger) than the actual estimate is also reported on the x-axis in Panels A and C (Panels B).

Online Appendix B: Consistent Testing for Stochastic Dominance

To begin, as discussed in the text, suppose S^F and S^M denote the empirical Bayes estimates for female and male judges, respectively. $\left\{s_i^F\right\}_{i=1}^{N_0}$ is a vector of N_0 observations of S^F; $\left\{s_i^M\right\}_{i=1}^{N_1}$ is an analogous vector of realizations of S^M. Let $F^F(s) = \Pr(S^F < s)$ represent the cumulative density function (CDF) of S^F; $F^M(s)$ is similarly defined for S^M. To test for first-order stochastic dominance (FSD), we utilize the following Kolmogorov-Smirnov test criteria:

$$d = \sqrt{\frac{N_0 * N_1}{N_0 + N_1}} \min \sup_{s \in S} [F^F(s) - F^M(s)]$$

where S denotes the union support of S^F and S^M and min is taken over $F^F - F^M$ and $F^M - F^F$, in effect performing two tests in order to leave no ambiguity between the equal and unrankabke cases. Following Linton et al. (2005), our consistent testing procedure calls for

1. Computing the empirical CDFs ($\widehat{F}(\cdot)$) over the support of S for s_k points ($k = 1...K, K$ is 500 in the application).

2. Computing the differences $d_1(s_k) = \widehat{F^F}(s_k) - \widehat{F^M}(s_k)$ and $d_2(s_k) = \widehat{F^M}(s_k) - \widehat{F^F}(s_k)$.

3. Obtaining $\widehat{d} = \sqrt{\frac{N_0 * N_1}{N_0 + N_1}} \min \left\{\max \left\{d_1\right\}, \max \left\{d_2\right\}\right\}$.

If $\widehat{d} \leq 0$ and $\max \left\{d_1\right\} < 0$, then S^F is observed to first-order dominate S^M; if $\widehat{d} \leq 0$ and $\max \left\{d_2\right\} < 0$, then the reverse is observed. If $\widehat{d} > 0$, then there is no ranking in the first-order sense.

Inference is conducted using a block bootstrapping procedure where resampling is done at the judge level. The bootstrap test statistic, $\widehat{d}^*$, is obtained by estimating a judge random effect model for each bootstrap replication of the original sample.[1] Under this resampling scheme, if $\Pr\left\{\widehat{d}^* \leq 0\right\}$ is large, say 0.90 or higher, and $\widehat{d} \leq 0$, we infer FSD to a desirable degree of confidence.

[1] The number of bootstrap replications is 500.