NBER WORKING PAPER SERIES

COMPARING APPLES TO ORANGES: DIFFERENCES IN WOMEN'S AND MEN'S INCARCERATION AND SENTENCING OUTCOMES

Kristin F. Butcher
Kyung H. Park
Anne Morrison Piehl

Working Paper 23079
http://www.nber.org/papers/w23079

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2017

This paper was prepared for a Festschrift for Bob LaLonde. We thank Bob LaLonde for his teaching, mentorship, and example. We have received many helpful comments from Festschrift participants, seminar participants at Amherst College, UVA Batten School, Southern Economic Association conference, and Wellesley College. We thank Hui Li, Morgan Matthews, Caitlin McCarey, and Emily Rothkin for outstanding research assistance. We are also grateful for Helen Pedigo, Scott Scultz, Kunlun Chang, and Brenda Harmon at the Kansas Sentencing Commission for their help with the sentencing data. Any errors or interpretations are our own and not of the Kansas Sentencing Commission. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Comparing Apples to Oranges: Differences in Women's and Men's Incarceration and Sentencing Outcomes

Kristin F. Butcher, Kyung H. Park, and Anne Morrison Piehl

NBER Working Paper No. 23079
January 2017
JEL No. J16,K14,K42

ABSTRACT

Using detailed administrative records, we find that, on average, women receive lighter sentences in comparison with men along both extensive and intensive margins. Using parametric and semi-parametric decomposition methods, roughly 30% of the gender differences in incarceration cannot be explained by the observed criminal characteristics of offense and offender. We also find evidence of considerable heterogeneity across judges in their treatment of female and male offenders. There is little evidence, however, that tastes for gender discrimination are driving the mean gender disparity or the variance in treatment between judges.

Kristin F. Butcher
Department of Economics
Wellesley College
106 Central Street
Wellesley, MA 02481
and NBER
kbutcher@wellesley.edu

Kyung H. Park
Department of Economics
Wellesley College
106 Central Street
Wellesley, MA 02481
kpark4@wellesley.edu

Anne Morrison Piehl
Department of Economics
Rutgers, The State University of New Jersey
New Jersey Hall
75 Hamilton Street
New Brunswick, NJ 08901-1248
and NBER
apiehl@economics.rutgers.edu

A online appendix is available at http://www.nber.org/data-appendix/w23079

1 Introduction

The incarceration of women has received attention from scholars and policy makers in recent years, in large part because the incarceration rate of women is rising more quickly than those of men (Carson and Golinelli (2012)). Although the incarceration rate of women continues to be less than a tenth of the rate for men, there is concern that incarceration of women, who are likely to be custodial parents (Mumola (1999)), may have broad adverse consequences for both their own future outcomes and those of their children. Further, incarceration is expensive and there is interest in whether it is being effectively deployed to incapacitate or deter the worst potential offenders (Travis et al. (2014)).

Whereas incarcerated men tend to come from socioeconomically disadvantaged backgrounds, the women tend to be even worse off (Harrison and Beck (2006)). The literature on incarcerated women documents poor outcomes for the women and for their children (see Poehlmann, Dallaire, Loper, and Shear (2010) for a review of the child development literature). A body of work by Robert LaLonde and his co-authors has examined long-term changes for women and their children after the women's incarceration by assembling linked administrative records. This research, using techniques that exploit *within-person* variation, finds remarkably little adverse causal impact of incarceration on women's employment (Cho and Lalonde (2008)), welfare dependency (Butcher and LaLonde (2013)), and children's test scores (Cho (2009)).

In this paper, we step back to examine how women and men are treated at the sentencing stage, investigating punishment differences between women and

men who have been convicted of felonies. We use detailed data on the universe of convicted felons in Kansas from 1998-2011 to investigate the differences between women's and men's penalties, adjusting for detailed observables about the elements of the offense, criminal history, and case facts. Trends in overall and female incarceration in Kansas are similar to national patterns. Other advantages include: Kansas has as representative a sentencing system as exists among the varied jurisdictions in the U.S., high quality data, and random assignment of cases to judges. These features are all discussed in detail later in the paper.

The raw punishment gap between women and men in incarceration and in sentence length conditional upon incarceration both indicate that women receive more lenient treatment. On average, women are 14 (drug crime) to 20 (non-drug crime) percentage points less likely to be incarcerated and receive 44 (non-drug crime) to 12 (drug crime) percent shorter sentences conditional on incarceration. Using linear regressions and conditioning on available observables about the case, women are 5 to 6 percentage points less likely to be incarcerated than men and receive 2 to 9 percent shorter sentences for non-drug and drug crimes, respectively. We also investigate the punishment gaps for non-drug crimes using semi-parametric techniques (DiNardo et al. (1996)) to examine the treatment of women who have the same observable characteristics as men, finding evidence that the same activity by men and women is treated quite differently in the criminal justice system even though Kansas is a state where judges follow guidelines to structure sentencing discretion.

Given a sizable unexplained portion of the gaps in punishment, we turn to

examining the role of judicial behavior in determining these gaps with a focus on the extensive margin of incarceration versus probation. We do this in two ways. First, we use an event-study analysis (Jacobson, LaLonde, and Sullivan (1993), Jacobson, LaLonde, and Sullivan (2003)) to trace out the impact of entry into or exit out of the Kansas court system by "harsh" or "lenient" judges on female incarceration.[1] The entry of a "lenient" judge reduces the probability of female incarceration by 5 percentage points, whereas the entry of a "harsh" judge increases the incarceration probability by 5 roughly percentage points for women.[2] These represent large percentage effects on incarceration probabilities for women given that the average women's incarceration rate for non-drug related crimes is 0.121.

Second, we compute standard measures of between-judge variation in female incarceration rates. This analysis uncovers substantial judicial heterogeneity in sentencing; for example, being assigned the 75th percentile judge rather than the 25th percentile judge increases the likelihood of incarceration for women by roughly 8 percentage points, which constitutes a 67% increase. Interestingly, there is slightly more variation across judges in women's incarceration rates in comparison with men's. The heterogeneity is not driven by gender-based case assignment nor is it the result of sampling error. Thus, while women receive more lenient sentences on average, there is still considerable *ex-ante* uncertainty in the probability of incarceration tied to the identity of the presiding judge.

[1] "Harsh" and "lenient" judges are defined by whether a judge's incarceration rate of non-person crimes committed by male offenders falls in the upper or lower quartile, respectively.

[2] We find that judicial exit has much less effect on incarceration probabilities for reasons we will discuss.

Multiple reasons for differential treatment of women are proposed in various literatures; women are less effective bargainers on their own behalf (Redlich and Shteynberg (2015)), have weak social ties conditional on being in the criminal justice system, and are often used as leverage in criminal cases against intimates (Starr (2012)). While we cannot test these alternatives without additional information, we can test the popular conception that judges are chivalrous (Spohn (1999)) towards women by applying a rank-order test of taste-based discrimination (Anwar and Fang (2006), Park (2015)). This approach examines whether the rank-order of judicial incarceration rates depends on the offender's gender. Our statistical test does not reject the null that judges do not have tastes for discrimination in favor of female defendants. This suggests that gender differences in unobservable characteristics, such as social ties and effective bargaining, may better explain differences in sentencing outcomes between men and women.

The findings in this paper help shed light on the previous literature on female incarceration and its impacts. The fact that there is a substantial punishment gap between women and men, conditioning on all the observables of the case, suggests that women who are incarcerated are more negatively selected than the men who are incarcerated. This interpretation supports the findings in LaLonde's (and his co-authors') work on incarcerated women which looks at outcomes for the same women (and their children) before and after incarceration and finds little adverse impact of incarceration.

The findings here also comport with recent literature that exploits variation in the probability of incarceration that arises due to random case-assignment

across heterogeneous judges (Aizer and Doyle (2015)). This source of variation might have different implications for the impact of incarceration on various life-outcomes because it would compare women who are relatively more and less negatively selected. Our finding that judges exhibit considerable heterogeneity in their sentencing towards women supports use of this variation to study the impact of female incarceration.

The paper is organized as follows. Section II provides an overview of sentencing in Kansas, including descriptive statistics. We then turn to understanding the female-male punishment gap in Section III. Section IV investigates judicial heterogeneity. Section V concludes.

2 Background

Institutional Details

Kansas employs sentencing guidelines to structure the punishments of criminal offenses.[3] Each felony is associated with a guideline sentence that recommends either prison or probation as well as the sentence length. Figure 1 shows that the guideline sentence is a function of the severity of the crime and the felon's criminal history.[4] In each cell of Figure 1, the three numbers are the minimum, expected, and maximum sentence length (in months). The grey

[3]More than 20 states use guidelines to structure criminal sentencing. See Stemen (2004) for an evaluation of how well the Kansas guidelines meet their objectives.

[4]In Kansas, the criminal severity level is the analogue to the base offense level in the Federal Sentencing Guidelines. As in the Federal Guidelines, the criminal severity level is determined by the crime of conviction. The Kansas Guidelines also allow for the possibility that some cases will involve special circumstances that could warrant a sentencing enhancement. However, unlike the Federal system, Kansan judges do not decide whether to adjust the base offense level. Thus, in Kansas, there is no technical distinction between the base and final offense level.

Figure 1: Kansas Sentencing Guidelines

Non-Drug Offenses

Severity Level	Criminal History A	B	C	D	E	F	G	H	I
	3+ Person Felonies	2 Person Felonies	1 Person & 1 Non-Person Felonies	1 Person Felony	3+ Non-Person Felonies	2 Non-Person Felonies	1 Non-Person Felony	2+ Misdemeanors	1 Misdemeanor or No Record
1 (Most Severe)	653/620/592	618/586/554	285/272/258	267/253/240	246/234/221	226/214/203	203/195/184	186/176/166	165/155/147
2	493/467/442	460/438/416	216/205/194	200/190/181	184/174/165	168/160/152	154/146/138	138/131/123	123/117/109
3	247/233/221	228/216/206	107/102/96	100/94/89	92/88/82	83/79/74	77/72/68	71/66/61	61/59/55
4	172/162/154	162/154/144	75/71/68	69/66/62	64/60/57	59/56/52	52/50/47	48/45/42	43/41/38
5	136/130/122	128/120/114	60/57/53	55/52/50	51/49/46	47/44/41	43/41/38	38/36/34	34/32/31
6	46/43/40	41/39/37	38/36/34	36/34/32	32/30/28	29/27/25	26/24/22	21/20/19	19/18/17
7	34/32/30	31/29/27	29/27/25	26/24/22	23/21/19	19/18/17	17/16/15	14/13/12	13/12/11
8	23/21/19	20/19/18	19/18/17	17/16/15	15/14/13	13/12/11	11/10/9	11/10/9	9/8/7
9	17/16/15	15/14/13	13/12/11	13/12/11	11/10/9	10/9/8	9/8/7	8/7/6	7/6/5
10 (Least Severe)	13/12/11	12/11/10	11/10/9	10/9/8	9/8/7	8/7/6	7/6/5	7/6/5	7/6/5

Drug Offenses	A	B	C	D	E	F	G	H	I
1 (Most Severe)	204/194/185	196/186/176	187/178/169	179/170/161	170/162/154	167/158/150	162/154/146	161/150/142	154/146/138
2	83/78/74	77/73/68	72/68/65	68/64/60	62/59/55	59/56/52	57/54/51	54/51/49	51/49/46
3	51/49/46	47/44/41	42/40/37	36/34/32	32/30/28	26/24/23	23/22/20	19/18/17	16/15/14
4 (Least Severe)	42/40/37	36/34/32	32/30/28	26/24/23	22/20/18	18/17/16	16/15/14	14/13/12	12/11/10

Notes: In the grey and clear boxes, the presumptive sentence is probation and prison, respectively. Dashed boxes are "Border Box" cells in which the judge can issue a non-prison sentence subject to the availability of an appropriate rehabilitation program. The numbers in each cell represent different sentence lengths in months. The low and high values represent the minimum and maximum sentence, respectively. The intermediate value is the recommended or presumptive sentence length.

boxes indicate that the guideline sentence is probation, whereas in the clear boxes, the presumptive sentence is prison.[5] For example, a defendant who is charged with theft of $100,000 (a severity level 5 crime) and who has 1 prior non-person felony should expect between 38 to 43 months in prison, regardless of gender.[6]

[5] The dashed boxes are "Border Box" cells. The presumptive sentence for these crimes is prison, but the judge can choose probation without the departure being subject to review.

[6] "Person" and "non-person" crime are formal legal terms used in Kansas. In the Kansas Sentencing Guidelines Desk Reference Manual (2014), it states that: "The 'person' designation generally refers to crimes that inflict, or could inflict, harm to another person. Examples of person crimes are robbery, rape, aggravated arson, and battery. The 'nonperson' designation generally refers to crimes committed that inflict, or could inflict, damage to property. Nonperson crimes also include offenses such as drug crimes, failure to appear, suspended driver's license, perjury, etc."

There are three primary ways in which gender disparities can arise under the sentencing guidelines. First, the guidelines provide judges with the discretion to issue longer or shorter sentences within cell. Second, judges can formally depart from the guideline sentence, on either extensive or intensive margins, based on mitigating or aggravating factors.[7] However, formal departures are subject to appeal and can be reversed.[8] Roughly 14% of cases are sentenced with a formal departure. Third, judges have more freedom to depart from the guideline sentence when the crime violates a special rule.[9] Approximately 28% of all cases involve special rule violations, which include committing a person felony with a firearm, aggravated battery against a law enforcement officer, committing crimes for the benefit of a street gang, persistent sex offenses, and more. The modal violation is committing a crime while on probation, parole, conditional release, post-release supervision, or felony bond (72%). In this case, the judge can order a prison sentence for even low severity crimes without being subject to formal review.[10]

Descriptive Statistics

Our data are the universe of convictions for felonies in Kansas from 1998 to

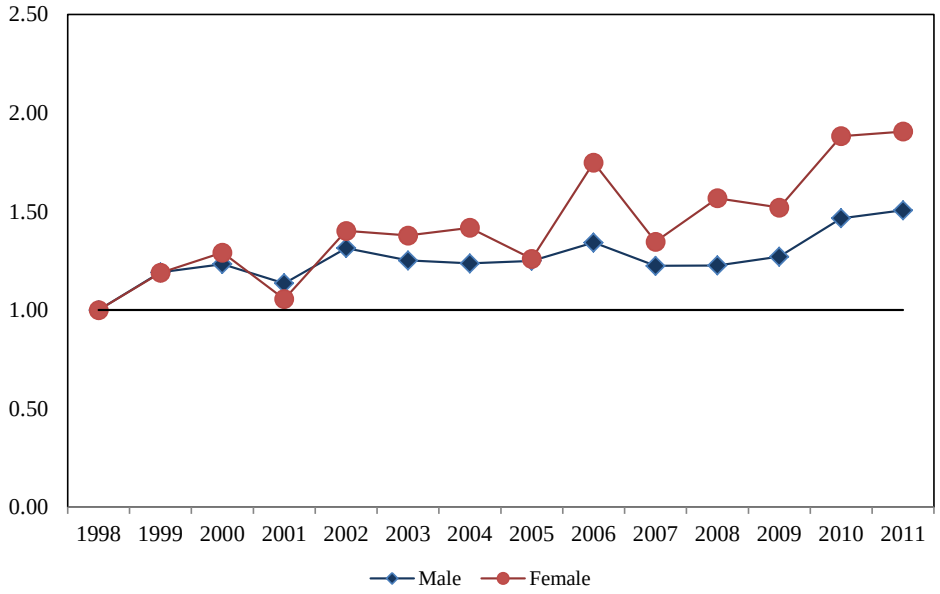
[7]Some examples of departing factors are whether the offender played a passive role in the crime, the crime is excessively brutal, the crime is a reaction to prolonged abuse, the crime is in self-defense, etc.

[8]There are limits to upward durational departures. The departure cannot exceed twice the base sentence length. In cases involving multiple convictions, the upward departure cannot exceed four times the base sentence length.

[9]As noted in footnote 4, this practice differs from that under Federal Guidelines. In Kansas, when special rule violations apply, judges are free to deviate from the sentence prescribed by the Guidelines without being subject to review.

[10]Note that the majority of criminal cases in Kansas, as elsewhere, are resolved by negotiated pleas. Judges are responsible for recording final outcomes, and for submitting justifications for departure.

Figure 2: New Court Commitments to Prison by Gender (Indexed to 1998)



Note: Calculations are based on data on the universe of convicted felons in the state of Kansas from 1998 to 2011.

2011.[11] Figure 2 shows data for women and men who are newly committed to prison by year indexed to 1998. The numbers are rising for both, but as Figure 2 demonstrates, the rise is steeper for women, with a 55 percent increase in incarceration for women over the period compared to a 30.5 percent increase for men.

Table 1 shows the demographic breakdown for this dataset which includes all persons who are newly sentenced for felony crimes (including both those who do and do not get sentenced to prison). Overall, about 19% of this population is female, 65% is white, 25% is black, and 10% is Hispanic. The last column shows the differences between women and men. The women are

[11]If an individual has multiple convictions over time, then he or she will be in the data multiple times. We do not have individual identfiers.

more likely to be white, and are about more than one and a half years older than the men, and all of these differences are statistically significant.

Table 1: Descriptive Statistics by Gender

	Overall	Female	Male	FM Diff
Female	0.188			
	(0.001)			
White	0.650	0.710	0.636	0.074
	(0.001)	(0.003)	(0.002)	(0.004)
Black	0.245	0.231	0.248	-0.017
	(0.001)	(0.003)	(0.001)	(0.003)
Hispanic	0.105	0.059	0.116	-0.057
	(0.001)	(0.002)	(0.001)	(0.002)
Age	30.926	32.254	30.619	1.635
	(0.030)	(0.068)	(0.033)	(0.076)
Age | No Prior Felonies	29.409	30.710	28.976	1.734
	(0.058)	(0.116)	(0.067)	(0.134)

Note: N = 119,081. Calculations are based on data on the universe of convicted felons in the state of Kansas from 1998 to 2011. The means are computed by regressing the variable on a constant and the mean gender difference by regressing the variable on a female indicator. Standard errors are in parentheses.

Table 2 shows differences in the types of crimes for which women and men are convicted. We have grouped crimes into drug and non-drug crimes, and indicated person crimes among the non-drug crimes. Kansas has separate sentencing grids for drug and non-drug crimes. While the main analyses examine non-drug crimes, we show the full set of offenses here and report some specifications for drug offenses where we can do so without disrupting the flow of the analysis. Panel A shows the fraction of women and men whose offenses fall into these categories. Sixty-three percent of the women are sentenced for non-drug crimes, and 68 percent of the men are; 33% of men are sentenced for a person crime, but only 15% of women are sentenced for person crimes.

Panel B of Table 2 lists the top 10 crimes for women separately for non-

Table 2: Types of Crime by Gender

Panel A: Gender Differences in Crime Type

	Female	Male
Person Crime	0.145	0.327
	(0.003)	(0.001)
Non-Drug Crime	0.631	0.682
	(0.003)	(0.002)

Panel B: Most Common Offenses Among Females

	% Conditional on Gender		
Top 10 Non-Drug Related Convictions for Females:	Female	Male	Female/Male
Forgery	27.88	6.79	4.11
Theft, between $1,000 and $25,000	15.44	10.64	1.45
Theft, less than $1,000	4.26	1.19	3.58
Burglary of a non-home	3.07	5.56	0.55
Making a false writing	2.98	1.02	2.92
Identity Theft	2.50	0.58	4.31
Burglary of a home	2.14	6.41	0.33
Giving a worthless check	2.12	0.72	2.94
Driving while a habitual violator	2.11	3.41	0.62
Aggravated battery, intentional bodily harm	1.80	3.80	0.47

	% Conditional on Gender		
Top 10 Drug Related Convictions for Females:	Female	Male	Female/Male
Possesion (1st offense)	57.67	45.27	1.27
Sale within 1,000 ft of school (Narcotics)	12.02	12.58	0.96
Sale within 1,000 ft of school (Hallucinogenic Drugs)	6.57	11.85	0.55
Possession (Hallucinogenic Drugs)	5.20	11.23	0.46
Drug, Sale of Opiates, Opium, or Narcotics	4.26	2.91	1.46
Use or Possession of Drug Paraphernalia	3.75	3.68	1.02
Unlawful Manufacture of Controlled Substance	2.41	4.04	0.60
Unlawful Possession of Ephedrine and etc.	2.13	1.61	1.32
Possession (2nd offense)	1.06	1.01	1.05
Drug Distribution (1st Offense)	1.06	0.75	1.41

Note: Calculations are based on data on the universe of convicted felons in the state of Kansas from 1998 to 2011.

drug and drug related offenses, ordered by their prevalence among women. The first two columns show the probability of the offense type conditional on female and male, respectively, and the third column shows the relative likelihood of the offense. Among non-drug related offenses, there are stark

differences in the types of criminal convictions between women and men. In particular, women's offenses are disproportionately concentrated among the types of criminal offenses that are associated with economic deprivation. While the top two categories, "forgery" and "theft between $1-$25K", account for 43% of convictions for women, they comprise only 17.4% of convictions for men. Moreover, the relative probabilities show that women are roughly three to four times more likely to face a criminal conviction of forgery, theft less than $1K, making a false writing, giving a worthless check, and identity theft in comparison with men. These statistics highlight the gender difference in the propensity to commit certain types of non-drug related crimes.

The top 10 drug crimes among women and men show more overlap in the categories and their rankings. For example, none of the ratios among the top 10 drug related offenses exceeds two. Nonetheless, there are still notable differences in the distribution of offenses between gender groups. The top two categories - "Possession (1st offense)" and "Sale within 1,000 ft of school (Narcotics)" - account for 70 percent of the women's convictions and 57 percent of the men's. In addition, a formal chi-square goodness of fit test for whether the distribution of drug crimes (and non-drug crimes) is the same across men and women is rejected at all conventional levels of statistical significance.

Table 3 shows the sentencing outcomes and case characteristics overall and by gender. Overall, 26% of individuals receive a term of incarceration, and the average sentence length conditional on incarceration is 43 months[12]; among

[12]In Kansas, the prison sentence is a strong predictor of actual time served. For example, for crimes committed on or after April 20, 1995, the maximum reduction of a prison term is 15%. For crimes committed on or after January 1, 2008, offenders who are sentenced to prison for low severity offenses can receive a maximum reduction of 20%.

those who receive probation, the average length is 19 months. There are large differences in men's and women's sentencing outcomes, with women being on average 18 percentage points less likely to receive a term of incarceration. Among those who are sentenced to prison, women's sentences are about 15 months shorter; among those on probation, women's sentences are about a month shorter.

The bottom panel of Table 3 shows the "Case Facts" as reported in the administrative records. These include factors about the crime that might enter into the sentencing guidelines or into the judicial decision. One can see that women frequently have more counts for a given charge than men, but the average severity of the charges is significantly lower. Women also have less serious criminal histories and are less likely to have committed a special rules violation. Women are less likely than men to have retained private counsel, a marker of having fewer resources with which to defend themselves, and are more likely to have entered a plea agreement and less likely to have objected to the official presentation of their criminal history.[13] They are more likely to have been released on felony bond prior to trial. Finally, women are about 3.7 percentage points (over 10 percent) more likely than men to have had a mental health or substance abuse evaluation. These differences point to women

[13]Even though over 95% of cases are resolved via plea, attorneys should endogenously incorporate judicial preferences and bargain in the "shadow of the judge" (LaCasse and Payne (1999)). In Kansas, attorneys know the identity of the presiding judge during plea negotiations and judges have power to veto plea agreements. One way to empirically assess the judge's influence is to see whether the residual variation in incarceration is more pronounced *within* versus *between* judicial districts since within-district variation holds fixed any district level characteristic including the sentencing preferences of the prosecutor. A simple decomposition shows that nearly 70% of the residual variation in incarceration that partials out the usual set of case facts is *within* district.

Table 3: Sentencing Outcomes and Case Characteristics by Gender

	Overall	Male	Female	F-M Diff
Sentencing Outcomes				
Incarceration	0.260	0.294	0.113	-0.181
	(0.001)	(0.001)	(0.003)	(0.003)
Prison Length (in Months)	42.942	44.203	28.771	-15.431
	(0.367)	(0.382)	(1.280)	(1.336)
Probation Length (in Months)	19.066	19.324	18.174	-1.149
	(0.026)	(0.029)	(0.055)	(0.062)
Case Characteristics				
Total Counts	1.269	1.262	1.299	0.037
	(0.003)	(0.004)	(0.008)	(0.009)
Severity (Non-Drug Crimes)	3.353	3.463	2.837	-0.626
	(0.007)	(0.007)	(0.016)	(0.018)
Severity (Drug Crimes)	1.448	1.470	1.367	-0.104
	(0.004)	(0.004)	(0.008)	(0.009)
Criminal History	3.886	4.111	2.911	-1.200
	(0.008)	(0.008)	(0.017)	(0.019)
Object to Criminal History	0.049	0.053	0.031	-0.023
	(0.001)	(0.001)	(0.001)	(0.002)
Person Crime	0.294	0.329	0.146	-0.183
	(0.001)	(0.001)	(0.003)	(0.003)
Non-Drug Crime	0.674	0.684	0.631	-0.052
	(0.001)	(0.002)	(0.003)	(0.003)
Special Rule Violations	0.277	0.285	0.245	-0.040
	(0.001)	(0.001)	(0.003)	(0.003)
Private Counsel	0.236	0.244	0.199	-0.045
	(0.001)	(0.001)	(0.003)	(0.003)
Released on Felony Bond	0.571	0.539	0.714	0.175
	(0.001)	(0.002)	(0.003)	(0.004)
Plea	0.956	0.953	0.973	0.020
	(0.001)	(0.001)	(0.001)	(0.002)
Mental, Alcohol, or Drug Evaluation	0.303	0.296	0.333	0.037
	(0.001)	(0.001)	(0.003)	(0.003)

Notes: N=119,081. Means and mean differences are computed by running regressions of a variable on a set of race indicators using observations without missing data. The scale for severity varies for non-drug vs. drug related crimes (1 to 10 for non-drug crimes and 1 to 4 for drug crimes). The scale has been inverted so that higher values are associated with more severe crimes. Criminal history is on a 1 to 9 scale with higher values reflecting more extensive prior records. Person crimes inflict physical or emotional harm to another person (e.g. robbery, rape, aggravated arson, and battery.)

as having less severe criminal histories and less severe offenses, while having fewer resources with which to defend themselves, at the same time having more mental and physical health issues with which to contend.

Figure 3 presents the joint distribution of severity and criminal history for men and women separately. Recall from Figure 1 that the intersection between criminal history and severity is what determines the recommended sentencing in those states, like Kansas, that adhere to sentencing guidelines. Women's distribution of criminal acts is much more crowded into the low-severity and low-criminal-history portion of the distribution. For men, while more crimes are in the low-criminal-history/severity portion of the distribution, the high-severity and high-criminal-history portion of the distribution is much more populated.

The descriptive statistics make clear that it is not surprising that women are less likely to be remanded to prison than men, and that conditional on prison, their sentences are lighter, since the measures of the severity of their crimes and their criminal histories are both lower. In what follows, we use a variety of techniques to compare incarceration and conditional sentence length between women and men who are similar along observable characteristics.

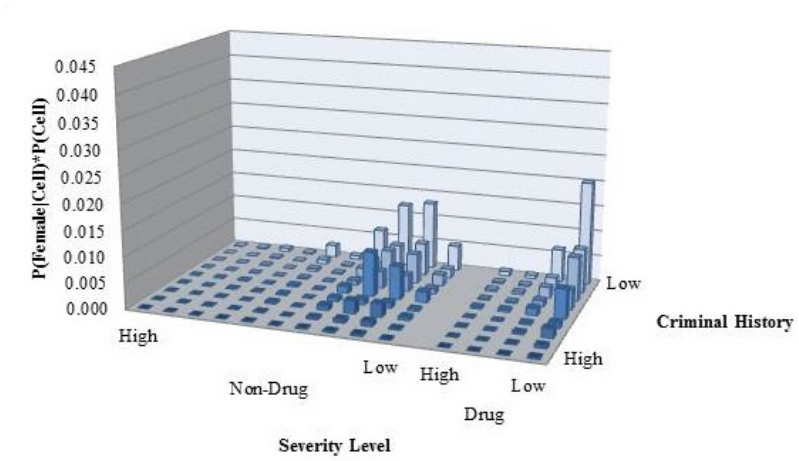
3 Female-Male Sentencing Gaps

3.1 Regression Analysis

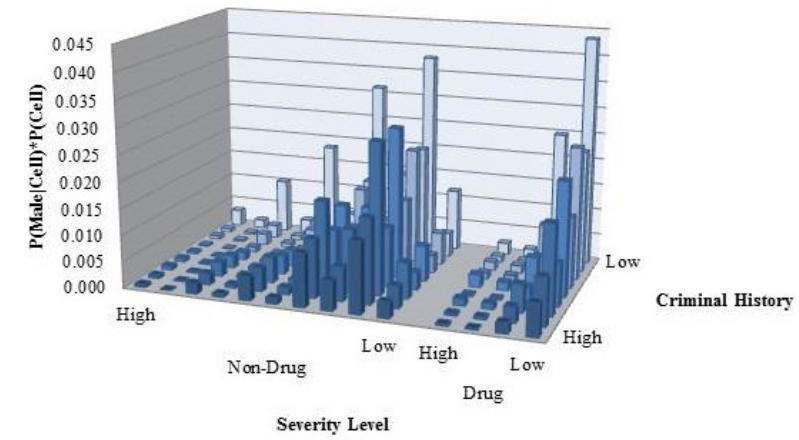
We begin our investigation using regression analysis and later pivot to semi-parametric re-weighting techniques in order to examine the extent to

Figure 3: Joint Distribution of Gender and Sentencing Cells

(a) Females



(b) Males



Note: The left side reflects the non-drug crime portion of the sentencing grid, whereas the right side is the drug crime portion of the sentencing grid.

which observed case facts explain the gender disparity in sentencing outcomes. The research literature that takes similar methodological approaches has been conducted on federal rather than state court cases (Starr (2015), Sorensen

et al. (2014), Mustard (2001)). This is a potentially salient difference given that federal crimes tend to involve far more serious offenses (along with a large number of drug cases, the seriousness of which is under contention (Sevigny and Caulkins (2004)) and are associated with more severe punishments. The types of men and women involved in federal cases are systematically different from those in state cases (Glaeser et al. (2000)). Thus, our findings provide evidence from a legal environment in which the gender gap receives disproportionately less attention, despite its arguably greater generalizablity.[14]

Our regression analysis focuses on two outcome variables: (i) whether or not the defendant receives a sentence of incarceration (=1, probation=0), and (ii) conditional on incarceration, the (log of) the sentence length:

$$Y_i = \alpha + \beta F_i + X_i\gamma + \epsilon_i \tag{1}$$

where Y_i denotes the outcome, F_i indicates that the offender is female, X_i represents a vector of case facts[15] and ϵ_i is the error term. The parameter of interest is β and represents the female-male sentencing disparity *ceteris paribus*. Its interpretation, however, can be ambiguous to the extent that some case facts themselves may be affected by gender discrimination.[16]

[14]While there is a fairly large economics of crime literature that uses state jurisdiction data, most studies focus on other determinants of sentencing besides gender. For example, Berdejo and Yuchtman (2013) and Lim (2013) use data from the state courts of Washington and Kansas, respectively, in order to study how judges respond to re-election concerns. More recent literature has focused on the effects of sentencing on labor market (Mueller-Smith (2014)) or marriage market outcomes (Charles and Luoh (2010)).

[15]Control variables are: the severity level of the offense, the offender's criminal history level, total counts, the offender's age and race, whether the defendant attains private counsel, committed a special rule violation, and resolved the case via plea.

[16]For example, in her study of federal jurisdiction, Starr (2015) finds larger gender incarceration gaps conditional on conviction. In her study, she is able to account for charge

The results are presented in Table 4. The top and bottom panels show estimates for non-drug and drug offenses, respectively. The first column presents the raw gaps in sentencing between women and men. The next column controls for indicators for the severity level of the crime. In most cases adding controls closes the size of the gap, although notably the prison length gap for drug crimes increases from 12 percent to 19 percent. When indicators for criminal history are included in column 3, the sentencing gaps fall dramatically. The incarceration gap closes to about 5 percentage points for non-drug and drug crimes; and the sentence length gap drops to about 2 percent for non-drug crimes and 7.6 percent for drug crimes.

In column 4, we replace indicators of the severity levels and criminal history with indicators for each sentencing cell. This specification is more flexible and accounts for any interactions among severity and criminal history. The estimated coefficients of interest are similar to those in the previous column. In the subsequent columns, we control for age and race/ethnicity, which although they do not enter into the sentencing grid, are correlated with recidivism and may influence judicial behavior; and as we saw in the descriptive statistics, women are older and more likely to be white than men. While these controls are statistically significant in the regressions, they do not materially affect the female punishment gap.

Finally, we add indicators for the case facts to the regression. These include indicators for whether the defendant is represented by private counsel, the plea status of the case, whether it was a person crime, the number of counts on

bargaining because her data includes information on the evolution of charges over the course of the plea negotiations.

Table 4: Regression Adjusted Gender Disparity in Sentencing Outcomes

				Panel A: Non-Drug Crimes			
Female-Male Gap in:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Incarceration Rates	-0.198***	-0.150***	-0.055***	-0.052***	-0.056***	-0.055***	-0.054***
	(0.006)	(0.009)	(0.008)	(0.008)	(0.007)	(0.007)	(0.005)
Log(Prison Length)	-0.442***	-0.168***	-0.019*	-0.030**	-0.033***	-0.031***	-0.024**
	(0.047)	(0.013)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
				Panel B: Drug Crimes			
Female-Male Gap in:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Incarceration Rates	-0.140***	-0.124***	-0.053***	-0.054***	-0.058***	-0.053***	-0.058***
	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)
Log(Prison Length)	-0.124***	-0.194***	-0.076***	-0.092***	-0.093***	-0.095***	-0.089***
	(0.031)	(0.019)	(0.023)	(0.022)	(0.022)	(0.022)	(0.022)
Covariates:							
Severity		Y					
Criminal History			Y				
Sentencing Cells				Y	Y	Y	Y
Age					Y	Y	Y
Race/Ethnicity						Y	Y
Case Facts							Y

Note: *** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1. Standard errors are in parentheses. Sample consists of 80,573 non-drug crimes and 38,958 drug related crimes. Among non-drug crimes, 22,842 cases are sentenced to prison, while among drug crimes, 8,175 are sentenced to prison. Case facts include indicators for whether the defendant is represented by private counsel, plea status, person crime, number of counts, and whether felon violates a special rule. Race/Ethnicity includes indicators for black and Hispanic. We control for age with indicators for 4 separate age groups, $< 25, 25-34, 35-44, 45+$. Sentencing cells include indicators for each cell in the sentencing grid. Severity includes indicators for each severity level, and criminal history includes indicators for each criminal history level. We also include separate indicators for values of covariates that are missing.

the charge, and whether the felon violates a special rule. Some of these are things that should affect the sentencing outcome, like whether there is a special rules violation (such as committing a crime while on felony bond), and others are things that we strongly suspect influence outcomes even if they are not included in statute (like private counsel). While many of these coefficients are statistically significant in explaining variation in incarceration and sentence length, they do not change the female punishment gap much.

Controlling for all the facts on the record for non-drug crimes, we still find

that women who are observably similar to men are 5.4 percentage points less likely to be incarcerated, and are sentenced to about 2% shorter sentences conditional on incarceration. For drug crimes, although the raw gaps (column 1) between women and men are smaller, controlling for observables does less to shrink them. Roughly 41% of incarceration and 71% of the sentencing gap remains unexplained for drug crimes.[17] Since drug and non-drug crimes require separate analysis due to the different sentencing grids, for the sake of brevity, we focus on non-drug crimes from this point forward. It is worth noting that non-drug offenses constitute the majority of crimes for both men and women.

3.2 Semi-Parametric Decompositions

While the regression analysis shows that case facts largely explain the mean difference in sentence length among non-drug crimes leaving only 5% of the original gap unexplained[18], it seems plausible that there could remain gender differences at other points in the distribution conditional on observables. In this section, we examine the actual and counterfactual distributions of prison terms conditional on incarceration using a semi-parametric re-weighting technique (DiNardo et al. (1996)). The counterfactual distribution for men is constructed by placing more weight on men who share case facts that are commonly observed among women. An analogous procedure can be used to construct the counterfactual distribution for women. This approach allows us

[17]These results are not driven by prison diversion programs for drug related offenses. Even when we restrict the sample to offenders who are not sentenced to this program, we get similar results. These estimates can be found in Section 3.7 of the Appendix.

[18]The raw gap for log of sentence length conditional on incarceration is -0.442 in column 1 and shrinks to -0.024 in column 7.

to examine the entire distribution in addition to evaluate the gender difference in means. It is worth noting that this approach abstracts from potential general equilibrium effects that could arise in the event that women and men actually commit the same types of crime.[19]

Figure 4 presents the actual and counterfactual distributions of sentence length conditional on incarceration for non-drug crimes. We focus on results that re-weight men to have the same covariate distribution as women because there are more men who have case characteristics similar to women in the data than there are women with similar characteristics to men as shown in Figure 3.[20] Thus, the exercise of re-weighting men has more support in the data than the exercise of re-weighting women to look like men.

Panel A shows how the distribution changes as we vary the weights using different sets of facts from the court records. Panel B graphs the differences between the two distributions and the vertical line indicates the modal value of the male distribution. If the distributions have the same mass at a given sentence length, then the graphs in Panel B will be at zero. If the line is below zero, there is more mass in the women's distribution at that point, and if the line is above zero, there is more mass in the men's distribution at that point.

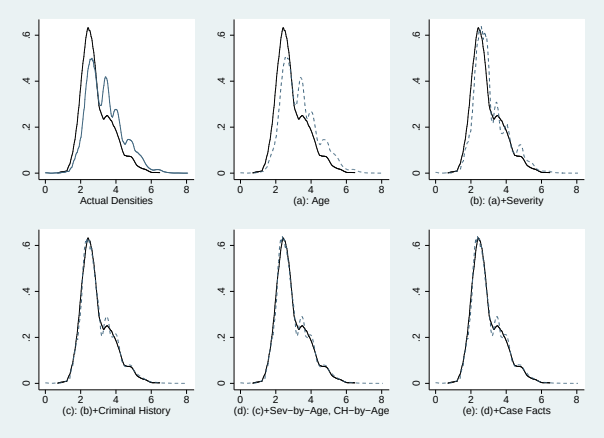
The first entry in Panel A shows the two actual distributions. The women's distribution is left-shifted compared to the men's. The women's distribution has a clear mode at about 3 log points, and the men's distribution has a number of spikes at about 3 and 4 log points. Each successive graph uses a different

[19]The details of the procedure are in Section 3.1 of the Appendix.

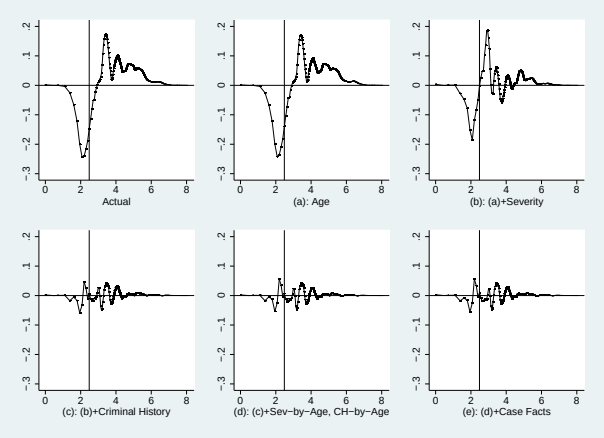
[20]Results are qualitatively similar when we conduct the exercise by re-weighting women to look like men (see Section 3.2 of the Appendix for these results).

Figure 4: Counterfactual Prison Length: Males Re-weighted

(A) Counterfactual Male and Actual Female Distributions



(B) Difference Between Distributions



Note: The counterfactual densities are estimated using kernel density estimation with the Epanechikov kernel, weighted by the estimated re-weighting function. The weighting function is estimated via probit models. Graph (a) controls for age, (b) adds the severity of the crime, (c) adds controls for criminal history, (d) adds interactions between age and criminal history and severity, and (e) adds the remaining case facts including race/ethnicity, type of counsel, whether a special rule applies, and counts.

set of observable characteristics in the underlying probit model to create the re-weighting function. The remaining graphs in Panel A demonstrate various

counterfactual distributions for the men, shown as dotted lines. The first of these is graph (a) which plots the counterfactual distribution for men if they were to have the same age distribution as women and the solid line is the actual distribution for women. Re-weighting men to have the same age distribution as women does not change the men's distribution of prison length much. When we add controls for severity, on the other hand, the men's counterfactual density looks much closer to the women's. Graph (c) adds controls for criminal history and the distributions are even more similar. Graph (d) adds interactions between age and criminal history and severity to the probit equation, and Graph (e) adds the remaining case facts including race/ethnicity, type of counsel, whether a special rule applies, and counts.

Turning to the graphs in Panel B, which show the differences in the distributions, we see that once the re-weighting function includes criminal history, the distributions stabilize and the differences are relatively close to zero.[21]

To summarize, our regression results show that even after conditioning on all of the observable criminal elements and despite the fact that Kansas uses sentencing guidelines that limit judicial discretion, about 30% of the gap in incarceration between men and women remains unexplained. Conditional on incarceration, gender gaps in sentencing length remain, but are small relative to the unexplained gaps in incarceration. Aside from the mean, the

[21]Note that we have also estimated gender difference in the mean sentencing outcomes among non-drug offenses using the re-weighting technique. The estimates shows that even after conditioning on all of the observable criminal elements and despite the fact that Kansas uses sentencing guidelines that limit judicial discretion, about 30% of the gap in mean incarceration between men and women remains unexplained. Conditional on incarceration, gender gaps in mean sentencing length remain, but are small relative to the unexplained gaps in incarceration. These results are qualitatively similar to our regression estimates and can be found in Section 3.2 of the Appendix.

re-weighting exercise shows little support of gender difference in other parts of the distribution of prison length conditional on incarceration. It is worth noting that the results thus far are mainly descriptive. Given that numerous actors could contribute to the unexplained gender disparity in incarceration, including judges, prosecutors, and police, we now turn our focus towards determining the role of judicial behavior in the incarceration probabilities of women.

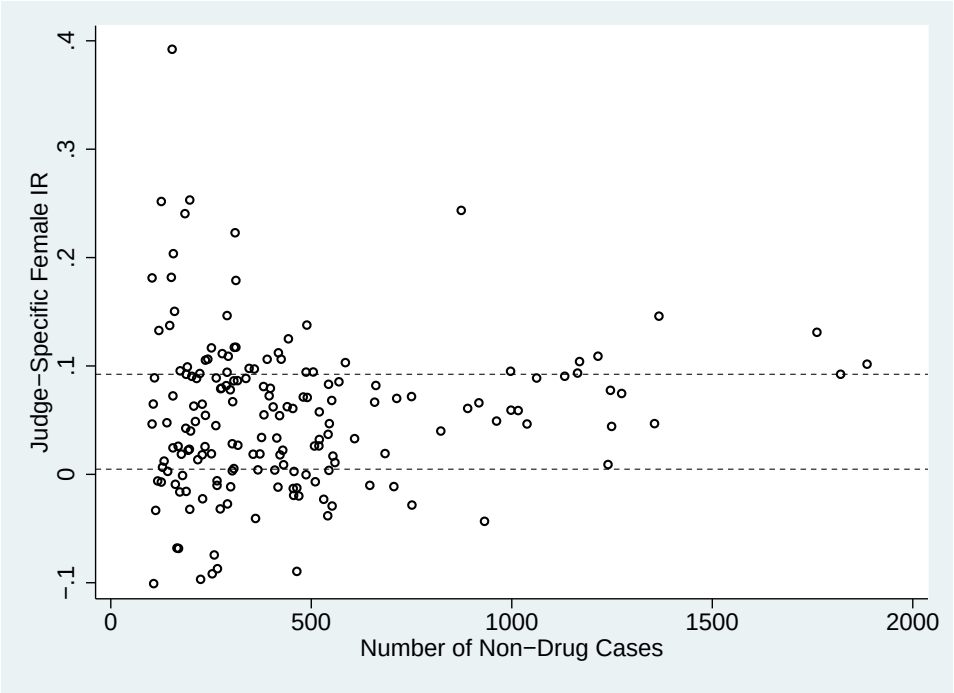
4 Judicial Heterogeneity

In light of structured sentencing and random case assignment, we might expect little judicial heterogeneity in sentencing outcomes. Figure 5 plots judge-specific female incarceration rates against total number of cases. The plot includes the 173 judges with more than 100 non-drug related cases. The two horizontal lines demarcate the 25th and 75th percentile judge. The plot shows considerable dispersion in judicial female incarceration rates. Being assigned the 75th as opposed to the 25th percentile judge increases the probability of incarceration by roughly 8 percentage points for women, which constitutes a 67% increase.

Figure 5 shows that judges who have seen fewer cases exhibit more dispersion but their estimates are also less informative due to sampling error. We will account for this in two ways. First, we will compute measures of dispersion that weigh judges by the inverse of the estimated variance of their fixed effect such that judicial incarceration rates that are estimated more precisely

receive more weight. Second, we will adjust estimates of the standard deviation associated with the distribution of judge fixed effects by subtracting off the mean of their estimated variances, which serves as a proxy for the sampling error variance (Aaronson, Barrow, and Sander (2007)). Even with these adjustments, the variation between judges is substantial.

Figure 5: Judicial Heterogeneity in Female Incarceration Rates



Note: Each dot is a judge's incarceration rate of female felons relative to the baseline judge. The baseline judge is normalized to 0. The weighted standard deviation is 0.064, where the weights are the inverse of the variance of the estimated judge fixed effect. The F-statistic associated with a joint significance test is 7.09. The judge fixed effects are estimated using variation orthogonal to the usual set of covariates. The horizontal lines denoth the 25th and 75th percentile judge, respectively. There are 173 judges total. Judges with fewer than 100 non-drug related cases are excluded. The sample is restricted to non-drug related crimes.

Before proceeding to quantify the extent of judicial heterogeneity, we consider the possibility that randomization is not achieved in practice. To test

whether cases are balanced across judges, we regress the presumptive sentence length on a set of judge fixed effects separately for each judicial district and then test whether or not the judge fixed effects are jointly equal to zero.[22] We also test for the presence of gender-based case assignment by including interactions between the offender's gender and judge fixed effects in the regression and then testing whether the interactions are jointly equal to zero. This will test whether or not some judges are assigned worse female offenders than other judges, on average.

The F-statistics and p-values associated with these tests are presented in Section 3.4 of the Appendix. On the whole, the evidence suggests that presumptive sentence length is balanced across judges in most judicial districts. In roughly 70% of the judicial districts, the p-value associated with the joint test of equality is above 0.05. There is also very little evidence of gender-based case assignment. The gender-by-judge fixed effects are jointly statistically significant at the 5% level in only one judicial district.

The fact that a few districts show imbalance further motivates our event-study analysis in which we exploit the entry into and exit out of districts by "harsh" and "lenient" judges. This approach mitigates residual concerns of non-random case assignment, since it is highly unlikely that case composition is endogeneously related with the timing of judicial entry or exit. Moreover, this approach utilizes *within-district* variation in sentencing, and thus accounts for all permanent unobserved district-specific factors that might influence the

[22]The presumptive sentence length is the middle value in the relevant cell of the sentencing grid. Given the multiplicity of observables and the 31 judicial districts, we focus on presumptive sentence length in order to provide a parsimonious and interpretable balancing test.

sentencing of female offenders; for example, prosecutorial or police behavior.[23]

4.1 Event-Study Analysis

In this section, we study the impact of entry or exit of "harsh" or "lenient" judges on female incarceration rates. Since we are focusing on incarceration for women, we define a "harsh" judge as one whose incarceration rate for male offenders who commit non-person crimes is in the top quartile and a "lenient" judge as one whose incarceration rate for male offenders who commit non-person crimes is in the bottom quartile of all judges.[24] To estimate the effects of entry/exit, we parameterize the regression model following Jacobson, LaLonde, and Sullivan (1993) and Jacobson, LaLonde, and Sullivan (2005):

$$y_{idt} = \gamma_t + \tau_d + \sum_{k=-2}^{3} D_{it}^k + X_i\beta + \epsilon_{idt} \tag{2}$$

where D_{it}^k are a set of timing of entry/exit fixed effects, X_i represents a vector of the usual set of case facts, γ_t denotes a set of year fixed effects, and τ_d are district fixed effects. The subscripts i denote the case, d denotes the district, and t reflects the year. The k superscript associated with the timing indicators denotes the k-th year relative to the event and takes values of $k = -2, -1, 0, 1, 2, 3+$. Districts that do not experience an event of entry or exit

[23]It is possible that prosecutors or other actors could change their behavior in response to changes in judicial composition. To assess this, we examine whether *predicted* incarceration (i.e. the fitted values from a regression of incarceration on case facts and other controls) changes with respect to judicial composition. We find no evidence supporting this hypothesis. These results are in Section 3.5 of the Appendix.

[24]We define "harsh" and "lenient" based on male incarceration rates in order to avoid the overfitting problem. Results are similar if we use combined incarceration rates, which suggests that judges treat female and male offenders similarly.

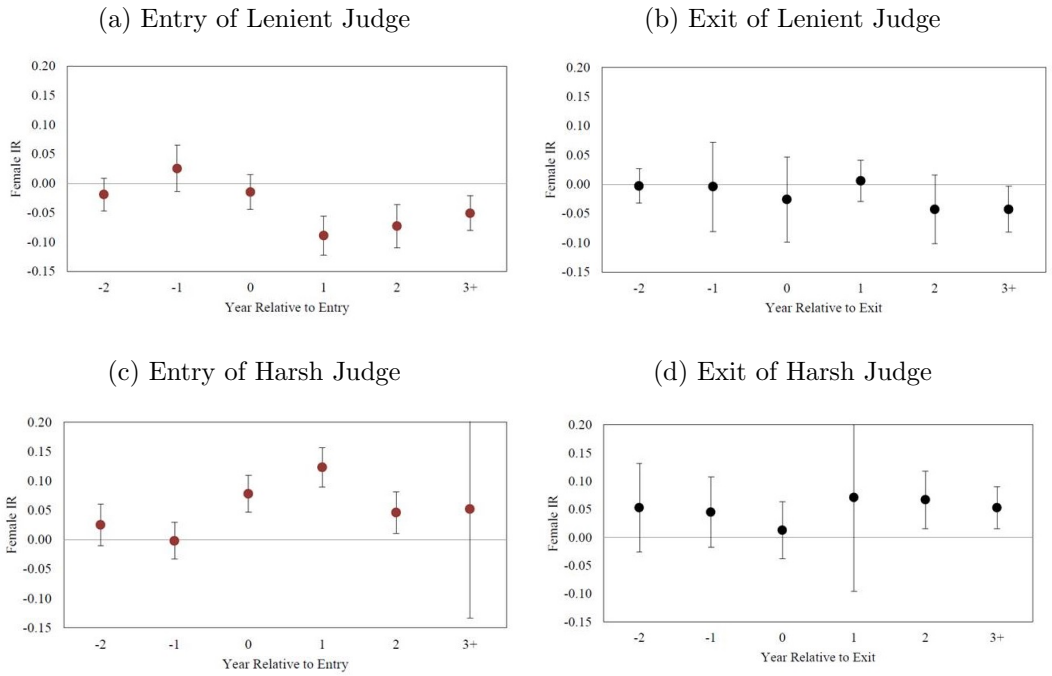
will identify the year fixed effects. Standard errors are clustered at the district-level.

Figure 6 presents four different plots which describe how female incarceration rates are affected by the timing of entry or exit. The left panel shows effects associated with entry of a lenient and then harsh judge, and the right panel shows effects associated with exit of a lenient and then harsh judge. Beginning with Panel a, we see that with respect to entry of a lenient judge, female incarceration rates decline not in the initial year of entry, but in subsequent years and that the decrease is persistent. Three years and beyond, female incarceration rates are roughly 5 percentage points lower in comparison with three years prior to the entry of the lenient judge. We also see no evidence of pre-trends prior to entry. The declines in response to entry of a lenient judge are statistically significant at the 5% level.

The entry of a harsh judge, Panel c, has opposite effects. In the year of entry of a harsh judge, female incarceration rates increase by nearly 10 percentage points. The magnitude is large given that the mean female incarceration rate is 11 percent. As with entry of a lenient judge, there is little evidence of a pre-trend in female incarceration rates. Three years and beyond, female incarceration rates are about 5 percentage points higher in comparison with the 3 years prior to entry, although this point estimate is not statistically significant.[25] On the whole, the time pattern is consistent with the narrative that the entry of a harsh judge leads to substantial increases in female incarceration.

[25]The standard error is large due to the fact that there are relatively few events and in the 17th judicial district, there is only 1 judge in the district leading to little within-district variation in the 3 years and beyond indicator.

Figure 6: Effect of Entry and Exit on Female Incarceration Rates



Note: "0" is the year of entry or exit. Harsh judge is defined as a judge whose incarceration rate of male non-person crimes is in the top quartile among all judges in the state. Lenient judges are in the bottom quartile in male non-person incarceration rates. Estimates are from a regression of incarceration on timing indicators, case facts, demographic characteristics, year effects, and district-fixed effects.

The right-hand panels show the effects associated with the exit of lenient and then harsh judges. Unlike entry, exit appears to have little effect on female incarceration rates. For example, in Panel b, most of the estimates associated with the exit of a lenient judge are close to zero and none are statistically significant at the 5% level. The exit of a harsh judge also has little impact on female incarceration rates. It is interesting that the effects of exit are not comparable to the effects associated with entry. Although it is beyond the scope of this paper, this suggests an important role for judicial learning

with respect to time in service. Recall that Figure 5 demonstrated that the variability in female incarceration rates across judges declines with the number of cases. Since the number of cases seen will increase with years in service, their sentences are less variable when judges exit.[26]

4.2 Gender Gaps and Judicial Heterogeneity

The previous section establishes that judicial entry has a substantial impact on the incarceration rate of women. In this section, we provide measures of dispersion associated with the distribution of judicial incarceration rates of both women and men. The judge-specific incarceration rates are covariate-adjusted for all of the observable characteristics of the case. Section 3.3 of the Appendix provides additional results from a re-weighting exercise that shows our measures of dispersion are robust to potential gender differences in case composition across judges. These analyses provide more general evidence on the importance of judicial heterogeneity in sentencing. In addition, these results allow us to explicitly compare judicial heterogeneity in sentencing of men versus women which speaks to the question of whether it is plausible to study the impact of incarceration on women's life-outcomes by exploiting variation associated with random case-assignment to heterogeneous judges.

[26]It is interesting that "harsh" judges affect sentences in their first year, but lenient judges only impact sentences after they have been on the bench for a year. One explanation may be related to the well-known fact that judges typically run election campaigns built on the promise that they will be "tough on crime" once elected into office (Gordon and Huber (2007), Berdejo and Yuchtman (2013)). Thus, it may not be surprising that harsh judges immediately respond with increases in female incarceration whereas lenient judges may exhibit more caution in issuing less punitive sentences until they learn how to navigate the criminal justice system. While we emphasize that this discussion is purely speculative, it is interesting that we observe other patterns in the data that are consistent with judges learning over time.

Table 5 shows estimates of the standard deviation, 75th vs. 25th percentile, and 90th vs. 10th percentile differentials associated with the distribution of judicial incarceration rates separately for women and men and that are both unweighted and weighted.[27] The unweighted estimates suggest that a 1 standard deviation change in the distribution of judges is associated with a 7.1 and 6.1 percentage point increase in the likelihood of incarceration for women and men, respectively. The weighted standard deviations for women and men are slightly smaller but still show substantial variation between judges. The weighted estimates imply that a 1 standard deviation change is associated with a 6.4 and 5.8 percentage point increase in the likelihood of incarceration for women and men, respectively. Given that the baseline female and male incarceration rates are 0.121 and 0.319, respectively, this constitutes a much larger effect size for women in comparison with men.

The adjusted standard deviation shows qualitatively similar results.[28] A 1 standard deviation change is associated with a 6.7 and 6.0 percentage point increase in the likelihood of incarceration for women and men, respectively. The 75/25 and 90/10 differentials also suggest that there is considerable heterogeneity across judges. To help put all of these statistics in perspective, the estimated incremental effects of obtaining private counsel, agreeing to a plea, and committing a person crime on the probability of incarceration are -3.9, -11.7, and 2.5 percentage points, respectively. Thus, judicial assignment impacts the probability of incarceration at a level comparable to that of key

[27]The weights are the inverse of the squared standard error of the judge fixed effect.

[28]The adjusted standard deviation subtracts off the mean of the estimated variances of the judge fixed effects.

case characteristics.

Table 5: Dispersion in Distribution of Judge Fixed Effects

	Unweighted			Weighted		
	Male	Female	FM Diff	Male	Female	FM Diff
Standard Deviation	0.061	0.071	0.010	0.058	0.064	0.006
	(0.003)	(0.005)	(0.004)	(0.003)	(0.005)	(0.003)
Adjusted Standard Deviation	0.060	0.067	0.007			
	(0.003)	(0.005)	(0.004)			
75/25 Difference	0.068	0.088	0.020	0.063	0.082	0.019
	(0.005)	(0.005)	(0.007)	(0.005)	(0.006)	(0.008)
90/10 Difference	0.152	0.148	-0.004	0.141	0.140	-0.001
	(0.011)	(0.011)	(0.011)	(0.012)	(0.010)	(0.012)

Notes: Bootstrapped standard errors are in parentheses. We estimate the judge fixed effects by regressing an indicator of incarceration on a vector of case facts and judge fixed effects. Rather than run the model separately by gender, we also include judge-by-gender interactions. The measures of dispersion correspond to the main judge fixed effects. The weighted statistics weight each judge fixed effect by the inverse of its estimated variance. The adjusted standard deviation subtracts off the mean of the estimated variances of the judge fixed effects from the unadjusted standard deviation. The sample is restricted to non-drug crimes.

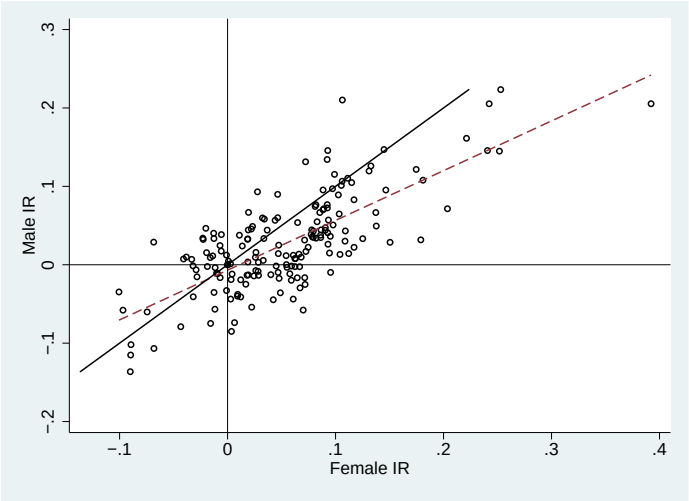
While these results show that judges exhibit considerable heterogeneity, the mechanisms that drive these differences are less clear. Perhaps the most straightforward explanation is that judges differ in punishment philosophies which translate into heterogeneity in sentencing. For example, some judges may have a higher willingness to afford an offender the opportunity to rehabilitate outside of prison. Other judges may have a stronger desire to incapacitate the same defendant. In Section 3.6 of the Appendix, we provide some descriptive evidence that is consistent with judicial differences in punishment philosophies. Specifically, we show that judges who incarcerate women at high rates also tend to incarcerate men who commit non-person offenses, low severity level offenses, and those with less extensive criminal histories. These patterns suggest that judges who are "tough on females" are "tough on property crime" more generally.

In addition, it seems plausible that statistical or taste-based discrimination could contribute to the observed judicial heterogeneity in the data. To assess this possibility, we develop a theoretical model of judicial sentencing (see Section 1 of the Appendix) that is a fairly straightforward adaptation of Anwar and Fang (2006) and Park (2015). Importantly, it allows for (i) judges to vary in their punishment philosophies, (ii) the prospect of statistical discrimination, and (iii) judges to have tastes for discrimination. An implication of the model is that in the absence of both forms of discrimination, judges should incarcerate both men and women at the same rate as long as the distribution of case facts is the same for men and women conditional on the observed elements of the case. While it is unlikely that we observe all the relevant variables in the judge's information set, it is worth reiterating that the data contains all of the case elements that determine the prescribed punishments by the state guidelines.

Figure 7 plots the regression-adjusted judicial incarceration rates of women and men along the x and y axes, respectively, as well as the 45^o line and the line of best fit. In the absence of discrimination, we should expect to see all of the judicial incarceration rates located close to the 45^o line with some deviation due to sampling error. The extent to which the line of best fit deviates from the 45^o line signals the potential for statistical or taste-based discrimination to affect judicial decisions. Interestingly, the plot shows that the line of best fit is considerably flatter than the 45^o line with a slope of 0.634.[29] A two-sided

[29]It is worth noting that the line of best fit deviates from the 45^o line with a flatter rather than steeper slope. As we illustrate in Section 1.2 of the Appendix, more punitive judges should have a higher willingness to incarcerate offenders who commit less severe offenses which are disproportionately associated with women. This implies that a more

Figure 7: Judicial Heterogeneity by Gender and Type of Crime



Note: Each dot represents judge-specific incarceration rates relative to the baseline judge, who is normalized to 0. The horizontal and vertical lines pass through the baseline judge. The incarceration rates are regression-adjusted for the usual set of covariates. There are 173 judges total. Judges with less than 100 cases are excluded. The sample is restricted to non-drug-related crimes. The solid black line is a 45^o line and the dashed red line is the OLS line of best fit through the judge specific incarceration rates.

hypothesis test of the null that the slope equals 1 is comfortably rejected with a t-statistic of -8.32.[30] Alternatively, a goodness of fit test of the null that judicial incarceration rates are the same for men and women is rejected with a p-value less than 0.001.

In summary, consistent with the event-study findings, we find that the identity of the presiding judge has sizable influence on the incarceration probability for both men and women. It is unlikely that this pattern is driven by systematic differences in case composition across judges. Not only are

punitive judge should exhibit a larger increase in the judicial incarceration rate of women in comparison with men. Thus, the fact that the line of best fit is flatter than the 45^o is not unexpected.

[30] A weighted regression that places more weight on judges whose female incarceration rates are estimated more precisely has negligible affect on this result.

cases randomly assigned, but in Section 3.3 of the Appendix, we also show that the observed judicial heterogeneity is robust to a re-weighting technique that equalizes the gender gap in case facts across judges. In addition, the misalignment between the line of best fit through the gender-specific judicial incarceration rates and the 45^o line implies that there is considerable judicial heterogeneity not only in the *levels* but also in the female-male *difference* in incarceration rates. This pattern hints at the possibility that perhaps statistical or taste-based discrimination influences judicial sentencing.

To identify which type of discrimination operates, we will apply the rank-order test of discrimination as in Anwar and Fang (2006) and Park (2015) which adapts the test to analyze judicial behavior. These studies forcefully argue that judicial heterogeneity in group-specific incarceration *rates* (in either levels or differences) cannot separately identify tastes for discrimination. Rather, if judicial incarceration *ranks* depend on gender, then this points to gender-based tastes for discrimination. Graphically, in the absence of tastes for discrimination, a plot of judicial incarceration *ranks* of men and women should lie atop the 45^o line. The next section presents the rank-order test.

4.3 Rank-Order Test of Taste-Based Discrimination

The key insight of the rank-order test is that in the absence of tastes for discrimination, more punitive judges should have higher male *and* higher female incarceration rates in comparison with other judges and thus judicial incarceration ranks should be the same regardless of the offender's gender. Table 6 provides an illustrative example of a rank-order violation. The table

shows gender-specific incarceration rates and ranks from the four judges in the Ford County district court.[31] We denote judge j's incarceration rate of gender g felons as γ_j^g where $g \in \{M, W\}$. Judge j's incarceration rank is provided in the parentheses such that higher ranks imply higher rates. The first cell indicates that judge A incarcerates men at the lowest rate ($\gamma_j^M = 0.227$) and women at the second lowest rate ($\gamma_j^M = 0.097$) among the four judges. On the whole, the table implies that judicial ranks depend on gender since the rank-order of men ($\gamma_A^M < \gamma_B^M < \gamma_C^M < \gamma_D^M$) differs from the rank-order of women ($\gamma_B^W < \gamma_A^W < \gamma_D^W < \gamma_C^W$).

Table 6: Gender-Specific Incarceration Rates and Ranks

Judge	γ_j^M	γ_j^W
A	0.227 (1)	0.097 (2)
B	0.256 (2)	0.074 (1)
C	0.262 (3)	0.170 (4)
D	0.304 (4)	0.146 (3)

Notes: The judicial incarceration rates are estimated using variation orthogonal to the usual set of case facts, including severity and criminal history. The judicial incarceration ranks are presented in parentheses. This example uses data from the 16th judicial district.

The open question is whether the rank-order violations are statistically significant or not. We conduct a statistical test in which the null hypothesis is that judges have no tastes for discrimination. We can formulate the null with

[31] Ford County is home of Dodge City, KS, the 14th most populous city in the state.

the following set of $\frac{k(k-1)}{2}$ inequality constraints:

$$H_o: \quad (\gamma_j^M - \gamma_{j'}^M)(\gamma_j^W - \gamma_{j'}^W) \geq 0 \ \forall j \neq j' \tag{3}$$

where k denotes the number of judges and, again, γ_j^M and γ_j^W represent the male and female incarceration rates for judge j, respectively. These inequalities convey the intuition that in the absence of taste-based discrimination, judge j should incarcerate *both* men and women at either higher or lower rates in comparison with judge j'. More extreme violations of these constraints will constitute stronger evidence of tastes for discrimination. Note that when this set of inequality constraints are satisfied a plot of judicial incarceration ranks should show data points that lie atop the 45^o line.

We compute a test statistic that is analogous to the F-statistic used in conventional tests in which the null hypothesis consists of equality constraints. Specifically, we compute the residual sum of squares from models that relax versus impose the inequality constraints in equation 3. Our test statistic, which we will denote as the $\bar{F}$-statistic, computes the difference in the residual variation across the two models, normalizes this by the residual variation from the unrestricted model, and adjusts for degrees of freedom. Intuitively, the $\bar{F}$-statistic will take larger values when the data exhibits inconsistency with the inequality constraints. On the other hand, if the data satisfy all of the inequality constraints, then the value of the $\bar{F}$-statistic will be zero. In this sense, the $\bar{F}$-statistic captures the change in model fit as we relax the inequality constraints.[32]

[32] An important difference between the F and $\bar{F}$-statistic is that the latter is distributed

There is an important consideration in statistical tests with inequality constraints that can be ignored in tests with only equality constraints. Because multiple parameter values can satisfy the inequality constraints, the asymptotic null distribution of our test statistic as well as the $(1-\alpha)$ quantile that serves as the critical value can vary depending on the location of the null. The crucial point made in the literature is that critical values are generally larger at locations where more inequality constraints bind. Thus, if the researcher could credibly test which constraints bind, then she could potentially leverage a more powerful statistical test by conducting inference under a re-centered null distribution where fewer inequality constraints bind. Our statistical procedure, which is detailed in Section 2 of the Appendix, adopts elements from well-developed techniques that allow one to select critical values in ways that increase statistical power while still controlling asymptotic size (Bugni (2010), Andrews and Soares (2010), Bugni et al. (2015), and Canay (2010)).

Table 7 shows results from the rank-order test. In particular, we show the district, the number of judges in the district, the $\bar{F}$-statistic, and its corresponding p-value. Note that we conduct the analysis separately for each district in order to control for district-level factors such as prosecutor or police effects. Again, values of the test statistic close to 0 indicate that the data are highly consistent with the inequality constraints whereas larger values constitute stronger evidence against the null. We find that for all but one district, the p-values sit comfortably above the 0.05 threshold required to

as a mixture of Snedecor's F-distributions due to the presence of inequality constraints (Wolak (1987)). While we cannot write closed form expressions due the large number of constraints, the distribution can be simulated via re-sampling techniques (Kudo (1963)).

achieve statistical significance. Only in the 19th district is the p-value below the conventional 0.05 level. On the whole, we cannot reject that the judicial incarceration *ranks* are independent of gender.

Table 7: Rank-Order Test Results

District	Number of Judges	$\bar{F}$ Statistic	P-value
1	3	0.007	0.690
2	4	0.082	0.538
3	14	0.104	0.467
4	3	0.056	0.641
5	4	0.124	0.558
6	4	0.188	0.455
7	4	0.110	0.482
8	6	0.070	0.717
9	4	0.099	0.572
10	11	0.025	0.942
11	6	0.030	0.876
12	2	0.840	0.104
13	5	0.166	0.500
14	6	0.000	0.998
15	2	0.002	0.656
16	4	0.010	0.874
17	1	N/A	N/A
18	30	0.034	0.972
19	3	2.327	0.005
20	4	0.319	0.237
21	3	0.438	0.187
22	2	0.000	0.975
23	3	0.068	0.533
24	2	0.004	0.484
25	4	0.136	0.491
26	5	0.007	0.936
27	5	0.281	0.343
28	6	0.000	0.998
29	15	0.064	0.802
30	3	0.275	0.345
31	5	0.108	0.566

Note: Our statistical procedure conducts the GMS pre-test procedures in order to determine which inequality constraints bind. We then simulate the empirical distribution of the test statistic under the re-centered null in order to compute p-values.

It is worth noting that the rank-order test plausibly identifies taste-based discrimination under a set of moderate assumptions. These are that (i) worse case facts signal that the offender poses a relatively high versus low threat to society, (ii) that judicial sentencing preferences are not substantially non-monotonic, and (iii) there is no gender gap in case facts across judges.[33] While the first seems intuitive and the third should hold due to random case assignment, recent research has shown that judges can exhibit considerable non-monotonicity in their sentencing preferences (Mueller-Smith (2014)). However, note that non-monotonicity should have the effect of *increasing* Type I error. For example, a judge who is particularly harsh on forgeries (the modal crime among women) will have higher female incarceration rates in comparison with other judges because of jurisprudence not gender-related tastes. Thus, our null finding is unlikely to be driven by non-monotonic sentencing preferences.[34]

While we find no statistical evidence of tastes for discrimination, it is important to emphasize that the rank-order test has low power. A rank-order violation implies tastes for discrimination, but as shown in Anwar and Fang (2006) as well as in Section 1 of the Appendix, the converse is not necessarily true. Thus, we cautiously interpret these results as providing suggestive evidence that tastes for discrimination are not a principle mechanism driving the earlier results on the mean gender difference or the variation in sentenc-

[33] A formal illustration of these assumptions is provided in Section 1.1 of the Appendix.

[34] If a researcher does find a rank-order violation in the data, then she should address these potential confounds more systematically. To account for non-monotonic sentencing preferences, a researcher could check whether the results are robust to "local" versions of the test in which attention is restricted to specific types of criminal offenses. To assess the importance of gender difference in case composition across judges, the researcher could apply a re-weighting technique that equalizes the gender gap in case facts across judges.

ing outcomes across judges. Instead, these results imply that the unexplained gender differences in sentencing are more likely driven by unobservables that differ across men and women, for example: the ability to negotiate favorable pleas, supportive social ties, or other relevant factors that are not captured in our analysis.

5 Conclusion

We find that men and women have very different crime profiles. Women are much more likely to commit non-violent, less severe crimes, and have less extensive criminal histories. However, conditional on all observable characteristics, both regression and semi-parametric results show that women are less likely to be incarcerated in comparison with similar men. In fact, roughly 30% of the gender gap in incarceration remains unexplained. We also find substantial heterogeneity in treatment of both female and male offenders by judges. Being assigned a judge one standard deviation above the mean judge will increase the probability of incarceration by 53% and 18% for women and men, respectively.

An oft-cited explanation for this "unexplained" gap and heterogeneity in sentencing is chivalry, or taste-based discrimination, in favor of female offenders by judges. We use a test of taste-based discrimination that relies on the rank-order of judicial incarceration rates of males and females and find that there is no statistical evidence of taste-based discrimination along gender lines. In other words, to a first approximation judges who are more lenient toward

women are also more lenient toward men. To the extent that there are violations in the rank-order of judicial gender-specific incarceration rates, we cannot reject that it is random.

Our results that there are large unexplained differences in incarceration rates for men and women who are observably similar that do not seem to be consistent with taste-based discrimination begs the question of what is driving the gender gap. If men and women differ systematically in terms of unobservable (to the researcher) characteristics that judges respond to, then that may result in gender gaps, but no rank-order violation in judicial behavior. For example, if judges do not like to incarcerate custodial parents, and women are more likely to be custodial parents than men, we may see substantial "unexplained" gender gaps. This suggests that other factors considered by the judge have to be worse for women than men in order for judges to get over the hurdle of incarceration for women. This is consistent with the finding, both here and elsewhere, that women who are incarcerated are more negatively selected than are men.

A larger remaining question is whether judges are making the right decision from a societal point of view. Are the gaps in punishment between men and women indicative of judges using the information they observe, but is not available to researchers in the data, to promote public safety in cost-effective ways? Without more information we cannot know. For now, we know from LaLonde (and co-authors) that the "types" of women that judges choose to incarcerate are the types for whom the prison term has little adverse effect on their employment, welfare receipt, and children's test score outcomes.

Whether that is consistent with reserving (costly) prison terms for only those who most threaten public safety is another question.

References

Aaronson, D., L. Barrow, and W. Sander (2007). Teachers and student achievement in the chicago public high schools. *Journal of Labor Economics 25*(1), 95–135.

Aizer, A. and J. J. Doyle (2015). Juvenile incarceration, human capital and future crime: Evidence from randomly-assigned judges. *The Quarterly Journal of Economics*, qjv003.

Andrews, D. W. and G. Soares (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica 78*(1), 119–157.

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

Berdejo, C. and N. Yuchtman (2013). Crime, punishment, and politics: an analysis of political cycles in criminal sentencing. *Review of Economics and Statistics 95*(3), 741–756.

Bugni, F. A. (2010). Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set. *Econometrica 78*(2), 735–753.

Bugni, F. A., I. A. Canay, and X. Shi (2015). Specification tests for partially identified models defined by moment inequalities. *Journal of Econometrics 185*(1), 259–282.

Butcher, K. F. and R. J. LaLonde (2013). Female offenders' use of social welfare programs before and after jail and prison: Does prison cause welfare dependency? *Working Paper*.

Canay, I. A. (2010). El inference for partially identified models: Large deviations optimality and bootstrap validity. *Journal of Econometrics 156*(2), 408–425.

Carson, E. A. and D. Golinelli (2012). Prisoners in 2012.

Charles, K. K. and M. C. Luoh (2010). Male incarceration, the marriage market, and female outcomes. *The Review of Economics and Statistics 92*(3), 614–627.

Cho, R. M. (2009). Impact of maternal imprisonment on children's probability of grade retention. *Journal of Urban Economics 65*(1), 11–23.

Cho, R. M. and R. J. Lalonde (2008). The impact of incarceration in state prison on the employment prospects of women. *Journal of Quantitative Criminology 24*(3), 267–267.

DiNardo, J., N. M. Fortin, and T. Lemieux (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica: Journal of the Econometric Society*, 1001–1044.

Glaeser, E. L., D. P. Kessler, and A. M. Piehl (2000). What do prosecutors maximize? an analysis of the federalization of drug crimes. *American Law and Economics Review 2*(2), 259–290.

Gordon, S. C. and G. Huber (2007). The effect of electoral competitiveness on incumbent behavior. *Quarterly Journal of Political Science 2*(2), 107–138.

Harrison, P. M. and A. J. Beck (2006). Prison and jail inmates at midyear 2005.

Jacobson, L., R. LaLonde, and D. G. Sullivan (2003). Should we teach old dogs new tricks? the impact of community college retraining on older displaced workers.

Jacobson, L., R. LaLonde, and D. G. Sullivan (2005). Estimating the returns to community college schooling for displaced workers. *Journal of Econometrics 125*(1), 271–304.

Jacobson, L. S., R. J. LaLonde, and D. G. Sullivan (1993). Earnings losses of displaced workers. *The American Economic Review*, 685–709.

Kudo, A. (1963). A multivariate analogue of the one-sided test. *Biometrika*, 403–418.

LaCasse, C. and A. A. Payne (1999). Federal sentencing guidelines and mandatory minimum sentences: Do defendants bargain in the shadow of the judge?*. *The Journal of Law and Economics 42*(S1), 245–270.

Lim, C. S. (2013). Preferences and incentives of appointed and elected public officials: Evidence from state trial court judges. *The American Economic Review 103*(4), 1360–1397.

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

Mumola, C. J. (1999). Incarcerated parents and their children. *children 37*, 44–6.

Mustard, D. B. (2001). Racial, ethnic, and gender disparities in sentencing: Evidence from the us federal courts. *Journal of Law and Economics 44*(1), 285–314.

Park, K. (2015). Do judges have tastes for discrimination? evidence from criminal courts. *Working Paper*.

Poehlmann, J., D. Dallaire, A. B. Loper, and L. D. Shear (2010). Children's contact with their incarcerated parents: research findings and recommendations. *American Psychologist 65*(6), 575.

Redlich, A. and R. Shteynberg (2015). Gender roles in juvenile and young adult plea decision-making.

Sevigny, E. L. and J. P. Caulkins (2004). Kingpins or mules: An analysis of drug offenders incarcerated in federal and state prisons. *Criminology & Public Policy 3*(3), 401–434.

Sorensen, T. A., S. Sarnikar, and R. L. Oaxaca (2014). Do you receive a lighter prison sentence because you are a woman or a white? an economic analysis of the federal criminal sentencing guidelines. *The BE Journal of Economic Analysis & Policy 14*(1), 1–54.

Spohn, C. (1999). Gender and sentencing of drug offenders: Is chivalry dead? *Criminal Justice Policy Review 9*(3-4), 365–399.

Starr, S. B. (2012). Estimating gender disparities in federal criminal cases. *University of Michigan Law and Economics Research Paper* (12-018).

Starr, S. B. (2015). Estimating gender disparities in federal criminal cases. *American Law and Economics Review 17*(1), 127–159.

Stemen, D. (2004). The Kansas Sentencing Guidelines: An Evaluation of the Proportionality of Sentences. Technical report, Vera Institute of Justice.

Travis, J., B. Western, and S. Redburn (2014). The growth of incarceration in the united states.

Wolak, F. A. (1987). An exact test for multiple inequality and equality constraints in the linear regression model. *Journal of the American Statistical Association 82*(399), 782–793.