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

Kristin F. Butcher Wellesley College
Kyung H. Park Wellesley College
Anne Morrison Piehl Rutgers University

1 Theoretical Framework

1.1 The Model

This section will provide a theoretical framework that will elucidate the key assumptions that underlie the rank-order test. The model is a simple adaptation of Anwar and Fang (2006) and Park (2015). The conceptual framework will provide more clarity on what precisely we can and cannot infer from judicial incarceration rates and ranks in regards to the role of statistical and/or taste-based discrimination in generating the observed patterns in data.

To begin, offenders are either $High$ or Low risk types in which $High$ types are more likely to re-offend and commit the types of crimes that are associated with high social costs. We denote risk type by $\tau \in \{H, L\}$. In addition, offenders vary by gender which we denote by $g \in \{M, W\}$ where M and W stand for men and women, respectively. The elements of a criminal case (e.g. the severity of the crime, the offender's criminal history, the type of counsel, and etc.) is indexed by θ. The distribution of θ, which we denote by $f_\tau^g(\theta)$, varies with τ and g and satisfies the monotone likelihood ratio property such that $\frac{f_H^g(\theta)}{f_L^g(\theta)}$ is strictly increasing in θ. In words, offenders who are associated with worse case facts have a higher relative likelihood of being a $High$ versus Low risk type.

Judges observe θ and choose whether to sentence an offender to prison or not. The judge does not, however, observe the offender's true type τ. If a judge "correctly" sentences a $High$ type to prison, then the payoff is 1. However, if a judge sentences a Low type to prison, then the judge incurs a cost, c_j^g, where j indexes the judge. We can motivate c_j^g several ways. For example, the judge may prefer to reserve incarceration for the worst type of criminals due to strong beliefs that Low types can better rehabilitate outside of prison. In this setup, judge j's expected net payoff to issuing a sentence of prison is $P(H|\theta) - P(L|\theta)c_j^g$. We normalize the expected net payoff to issuing a sentence of probation to 0. This is innocuous since the difference and not the level of expected utility affects choice. Applying Bayes Rule, it is straightforward to show that the judge's decision hinges on whether θ is below or above a threshold, θ_j^{g*}, which is determined by the following expression:

$$\frac{f_H^g(\theta_j^{g*})\pi^H}{f_L^g(\theta_j^{g*})\pi^L} = c_j^g$$

where π^τ denotes prior beliefs on τ. We now have the necessary elements to formally define taste-based and statistical discrimination and illustrate the

underlying mechanics of the rank-order test. We will use graphical illustrations to help visualize the key points.

Rank-Order Test of Taste-Based Discrimination ‘

In the model, judges are defined has having tastes for discrimination when $c_j^M \neq c_j^W$. Figure A1 will illustrate how tastes for discrimination can lead to a rank-order violation in judicial incarceration rates. The plot shows that judge j does not have tastes for discrimination since $c_j^M = c_j^W$ and thus, the marginal male and female offender would receive the exact same sentence under judge j. In contrast, judge j' is chivalrous towards women since $c_{j'}^W > c_{j'}^M$. Notice that because the judge's decision rule is to sentence an offender to prison whenever $\theta > \theta_j^{g*}$, the ordering of the judicial incarceration rates is not the same across gender since $\gamma_{j'}^M > \gamma_j^M$ but $\gamma_{j'}^F < \gamma_j^F$ where γ_j^g denotes the incarceration rate of gender g offenders for judge j. It is straightforward to show that when judges do not have tastes for discrimination, then a rank-order violation cannot arise. This is true even if judges were to engage in statistical discrimination due to the informative signal that gender may provide regarding the offender's risk type.[1] This motivates the rank-order test. If the rank-order of judicial incarceration rates depends on gender, then this implies judges engage in taste-based discrimination.

Key Assumptions of the Rank-Order Test

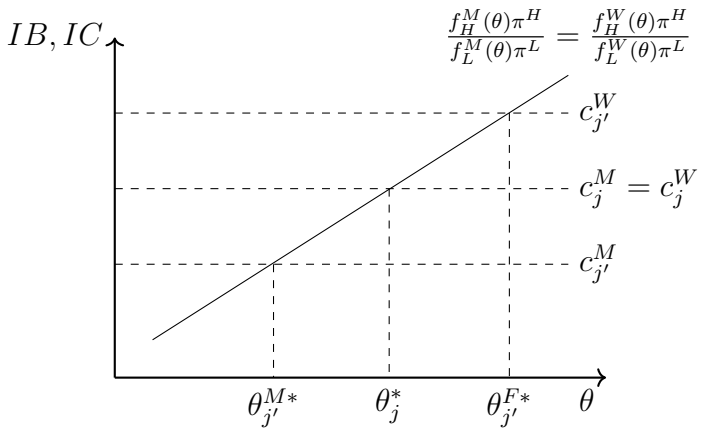
We will now briefly discuss the key assumptions.

- The MLRP plays an important role. Without it, the relative likelihood of being a $High$ versus a Low risk type is not necessarily a strictly increasing function of θ. In this case, the judge's legal standard is not uniquely determined and a rank-order violation could arise even in the absence of tastes for discrimination.

- The assumption that the judicial costs of incarceration, c_j^g, is independent of θ is important. Suppose, for example, that judge j prefers to issue severe sentences for forgery, an offense disproportionately committed by women, whereas judge j' prefers lenient ones. In this case, judge j may incarcerate women at higher rates than judge j' and men at lower rates than judge j' because of different sentencing preferences across crime type rather than taste-based discrimination. If judges exhibit considerable non-monotonicity in sentencing preferences, then it is possible for a rank-order violation to arise even in the absence of tastes for discrimination.

[1] Formally, we can say that judges statistically discriminate when $\frac{f_H^M(\theta)\pi^H}{f_L^M(\theta)\pi^L} \neq \frac{f_H^W(\theta)\pi^H}{f_L^W(\theta)\pi^L}$.

Figure A1: Rank-Order Test of Taste-based Discrimination



Notes: IB and IC stand for incremental benefit and cost of incarceration, respectively. c_j^g reflects the cost that judge j incurs when sentencing a Low type to prison. This cost is allowed to depend on the offender's gender. Taste-based discrimination is defined as $c_j^M \neq c_j^W$. θ_j^{g*} denotes the threshold that determines the judge's sentence. Finally, $\frac{f_H^g(\theta)\pi^H}{f_L^g(\theta)\pi^L}$ represents the relative likelihood that the offender is a $High$ versus Low type. It can be interpreted as the incremental benefit of sentencing an offender to prison.

- The gender difference in case composition across judges is important. The rank-order of judicial incarceration rates will depend on gender if some judges receive cases involving the $High$ type women and Low type men and others receive cases involving Low type women and $High$ type men.

The first assumption (e.g. the MLRP) seems reasonable to a first approximation. The other two assumptions may not necessarily hold. In particular, recent research shows that judges have non-monotonic sentencing preferences which casts some doubt on the assumption that judicial cost functions are independent of θ (Mueller-Smith (2014)). However, notice that relaxing these two assumptions should increase the likelihood of Type I error. The fact that we cannot reject the null of no taste-based discrimination in our data implies that these two forces are not sufficiently strong to generate a false rejection.[2]

Finally, while this has been pointed out in existing literature, we should

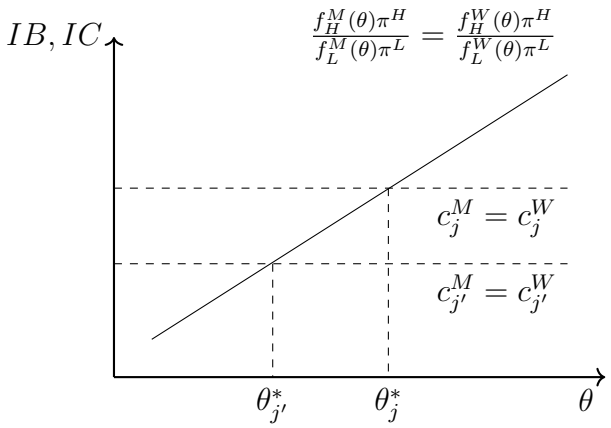
[2]If a researcher does find a rank-order violation in the data, then she should address these potential confounds more systematically. To account for non-monotonic sentencing preferences, a researcher could check whether the results are robust to "local" versions of the test in which attention is restricted to specific types of criminal offenses.

note that the rank-order test is a fairly conservative test of taste-based discrimination. There are scenarios in which some or even all judges could have tastes for discrimination but the rank-order of judicial incarceration rates will not depend on gender. Consider, for example, the following ordering of judicial costs $c_j^W > c_j^M > c_{j'}^W > c_{j'}^M$ which would imply the following ordering in judicial incarceration rates $\gamma_j^W < \gamma_j^M < \gamma_{j'}^W < \gamma_{j'}^M$. In this case, judge j' incarcerates men and women at higher rates than judge j. Thus, our results should be interpreted cautiously given that the rank-order test is a low-powered test of discrimination.

1.2 Judicial Incarceration Rates

In this section, we would like to formally show why judicial incarceration rates should all lie along the 45^o line in the absence of gender discrimination (both statistical and taste based). We will illustrate this point in the context of the model outlined above.

Figure A2: Judicial Incarceration Rates in Absence of Discrimination



Notes: IB and IC stand for incremental benefit and cost of incarceration, respectively. In this example, neither judge has tastes for discrimination since $c_j^M = c_j^W$ for both j and j' and there is no incentive to statistically discriminate since $\frac{f_H^M(\theta)\pi^H}{f_L^M(\theta)\pi^L} = \frac{f_H^W(\theta)\pi^H}{f_L^W(\theta)\pi^L}$.

Figure A2 shows a scenario in which neither judge has tastes for discrimination, since $c_j^M = c_j^W$ for both judges, nor will any judge engage in statistical discrimination as $\frac{f_H^M(\theta)\pi^H}{f_L^M(\theta)\pi^L} = \frac{f_H^W(\theta)\pi^H}{f_L^W(\theta)\pi^L}$. It follows that neither the decision rule of judge j nor j' will depend on the offender's gender even though each judge

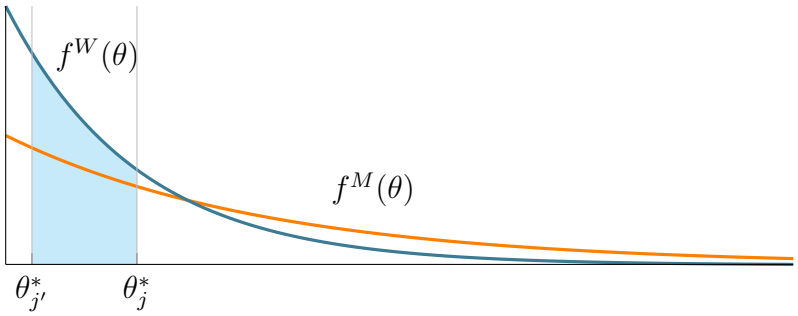
employs a different decision rule from the other. If we assume that the distributions of θ are equalized once we condition on the observed elements of the case, then the model predicts that any given judge must incarcerate men and women at the same rate such that each dot in Figure 7 in the paper aligns perfectly along the 45^o in the absence of discrimination. Systematic deviations from the 45^o will be a tip-off that judges engage in either statistical or taste-based discrimination.

As noted in the paper, the line of best fit deviates from the 45^o line with a flatter rather than steeper slope. We suggested that one possible explanation is that the distribution of case facts may differ with respect to gender even after conditioning on observable characteristics. To illustrate this point more formally, we present Figure A3 which highlights the expected changes in incarceration rates by gender as we move from a less to a more punitive judge. The plot shows blue and orange lines which reflect the hypothetical distributions of θ for women and men, respectively. The distributions are drawn such that women are more likely to be associated with lower values of θ, which is consistent with our descriptive evidence that shows women commit less severe crimes, on average. Judge j' is more punitive relative to judge j since $\theta^*_{j'} < \theta^*_j$. Importantly, the area shaded under the blue and orange curves between $\theta^*_{j'}$ and θ^*_j represent the difference in the women's and men's incarceration rate between the two judges, respectively. In this case, a comparison of the two judges shows that the increase in the judicial incarceration rate of women is smaller than the increase in the judicial incarceration rate of men because more punitive judges have a higher willingness to incarcerate less severe offenses which are disproportionately represented by women. Thus, in this framework, the fact that in Figure 7 in the paper the line of best fit is flatter than the 45^o is not necessarily unexpected.

1.3 Robustness of Rank-Order Test to Spillover Effects

The rank-test is robust to herding and/or spillovers across judges. We will illustrate this point is in the context of our conceptual framework. Panel (a) of Figure A4 shows a baseline scenario in which judges have incentive to set gender-specific decision rules because for a given θ, the relative likelihood of being a $High$ risk type is higher for men than for women (e.g. judges statistically discriminate against men). However, neither judge has tastes for discrimination since the incarceration costs are the same across gender for any given judge. The optimal decision rule dictates that a judge will sentence an offender to prison if θ exceeds the judicial threshold θ^{g*}_j and to probation otherwise. It follows that in the absence of tastes for discrimination, the

Figure A3: Understanding the Slope Less Than 1



Notes: The blue and orange lines represent the distribution of θ for women and men, respectively. The figure shows that as we move from judge j to j', there is a larger increase in the incarceration rate for women than for men since more punitive judges have a higher willingness to sentence offenders to prison for less severe crimes and women are overrepresented in these crime types.

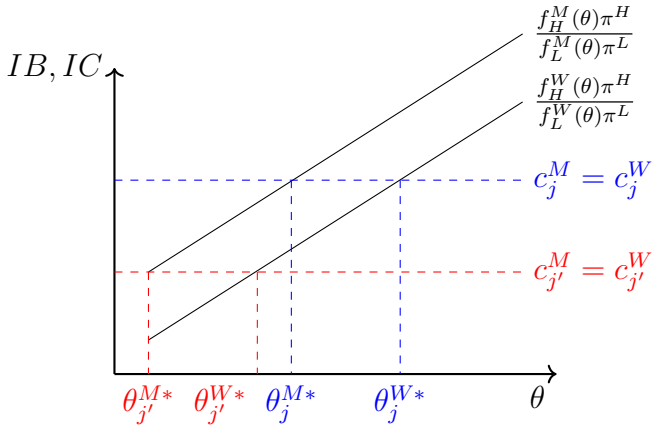
ranking of judicial incarceration rates will be the same for both female and male offenders. This is the key idea underlying the rank-order test.

Consider a thought experiment in which the entry of a harsh judge represents a shock that plausibly affects the sentencing preferences of others. To what extent do these spillover effects unwind the rank-order test? Suppose the entry of a harsh judge (whose cost function is not shown) has the effect of shifting the cost function of judge j' up or down. Panel (b) of Figure A4 illustrates upward shifts in the incarceration costs of judge j' from $c_{j'}$ to $\hat{c}_{j'}$ and then from $\hat{c}_{j'}$ to $\tilde{c}_{j'}$. Each dot is located at an intersection that determines the judge's threshold rule, θ^{g*}_j. For example, the x-coordinates of Points A and B reflect the decision thresholds for judge j' used to sentence men and women, respectively. In general, the dots that lie along the $\frac{f^M_H(\theta)\pi^H}{f^M_L(\theta)\pi^L}$ and $\frac{f^W_H(\theta)\pi^H}{f^W_L(\theta)\pi^L}$ lines represent the decision thresholds applied towards male and female offenders, respectively. Note that the increases in the incarceration costs of judge j' are associated with right shifts in her decision thresholds. This can be interpreted as judge j' becoming more lenient in response to the entry of a harsh judge.

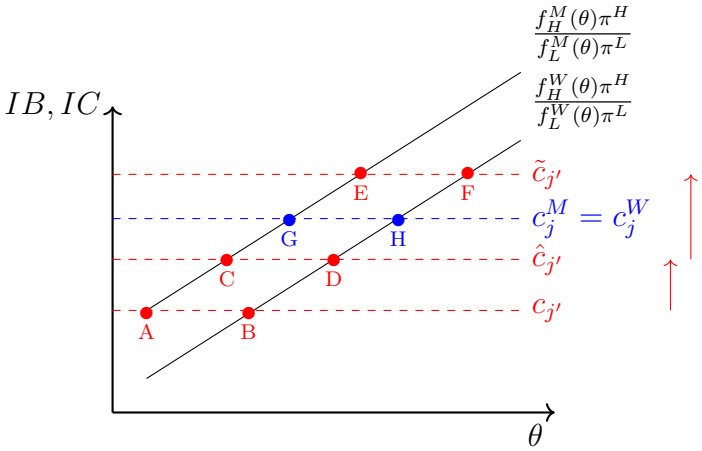
The key feature of this plot is that at each location of $c_{j'}$, the ordering of the judicial thresholds remains independent of the offender's gender. At baseline, the judicial thresholds are ordered such that $A < G$ and $B < H$. Given the judge's decision rule, this implies that judge j' will incarcerate men and women at higher rates in comparison with judge j. The same is true at

Figure A4: Robustness to Spillovers

Panel (a): Ordering of Judicial Thresholds at Baseline



Panel (b): Ordering of Thresholds With Spillovers



Notes: IB and IC stand for incremental benefit and cost of incarceration, respectively. This graph shows shifts in the incarceration costs of judge j' presumably due to the entry of a harsh judge. In this example, judge j becomes more lenient but the results do not hinge on this assumption. The x-coordinate of each dot represents the judge's decision threshold, θ_j^{g*}. Thus, as judge j' becomes more lenient, her decision thresholds shift to the right.

$\hat{c}_{j'}$. At $\tilde{c}_{j'}$, judge j' is now less punitive than judge j since $\tilde{c}_{j'} > c_j$. The

decision thresholds are ordered such that $G < E$ and $H < F$ which implies that judge j' will now incarcerate men and women at lower rather than higher rates than judge j. While the ordering of the judicial incarceration rates has switched, the crucial feature is that the ordering remains *independent* of the offender's gender. Even in the presence of spillover effects, the ordering of judicial incarceration rates will not depend on the offender's gender in the absence of tastes for discrimination. The robustness of the rank-order test highlights an advantage of an empirical test based on judicial incarceration *ranks* not *rates*.

2 Statistical Procedure

To implement the rank-order test, we adapt the procedure developed in Park (2015) to apply to the study of gender discrimination. The rank-order test requires non-standard statistical techniques. This is because the null hypothesis of no tastes for discrimination involves the following set of $\frac{k(k-1)}{2}$ inequality constraints:

$$H_o: \quad (\gamma_j^M - \gamma_{j'}^M)(\gamma_j^W - \gamma_{j'}^W) \geq 0 \ \forall j \neq j' \tag{1}$$

where k denotes the number of judges and γ_j^M and γ_j^W represent the male and female incarceration rates for judge j, respectively. This null hypothesis highlights the intuition that in the absence of taste-based discrimination, judges should exhibit consistency in sentencing; that is, 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.

The presence of inequality constraints in the null raises a concern. Because the null hypothesis specifies the *ordering* rather than the *level* of judicial incarceration rates, multiple parameter values can satisfy the inequality constraints under the null. In consequence, 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. A burgeoning econometrics literature on statistical inference in partially identified models offers well-developed techniques that allow the researcher 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)). While a formal and detailed treatment is beyond the scope of this paper, we would like to provide a simple example in $\mathbb{R}^2$ in order to illustrate key concepts.

Consider a data generating process $Y \sim N(\gamma, I)$ and suppose that we are interested in conducting the following statistical test:

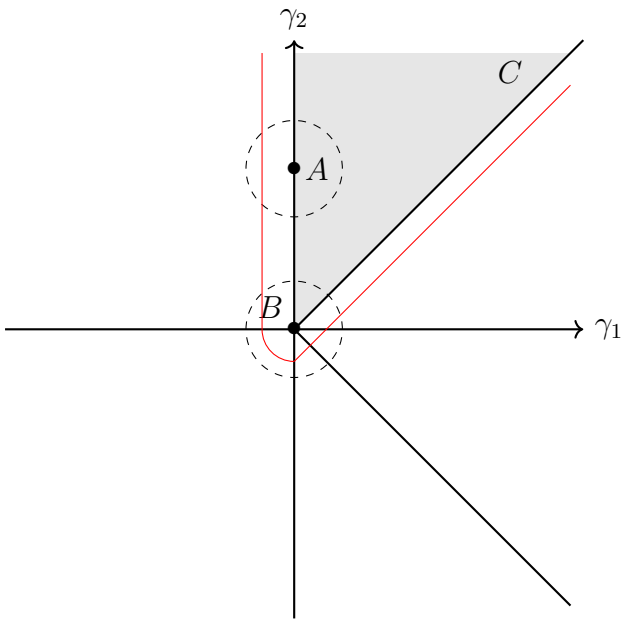
$$H_0 : \gamma_1 \geq 0, \gamma_2 \geq \gamma_1 \ \textit{vs.} \ H_1 : (\gamma_1, \gamma_2) \in \mathbb{R}^2 \tag{2}$$

Figure A5 will help visualize the complication that can arise in statistical tests with inequality constraints. The grey shaded area represents the cone, C, which contains all the parameter values that satisfy the inequality constraints imposed under the null. The red line denotes the set of points y that are all some distance l away from the cone and the dashed black lines are the contour lines associated with each data process. Point A represents the location of a data process at which only one constraint binds (e.g. $\gamma_1 = 0$ but $\gamma_2 > \gamma_1$) whereas at Point B both constraints bind (e.g. $\gamma_1 = 0$ and $\gamma_2 = \gamma_1$). While Points A and B both represent locations that are consistent with the null hypothesis, notice that the tail probabilities (e.g. the area of the circles that lie beyond the red line) associated with these two null distributions are noticeably different.[3] In particular, the $(1-\alpha)$ quantile of the null distribution centered at B is larger than its counterpart at A. In general, it is true that the critical values are larger at points in the parameter space where more inequality constraints bind.

The crucial point made in the literature is as follows: If the researcher knew which constraints were actually binding, then she could potentially leverage a more powerful statistical test by conducting inference under a re-centered null distribution where fewer inequality constraints bind. Recent literature has made considerable progress on developing "pre-test" procedures that estimate slackness parameters in order to determine which inequality constraints bind (Andrews and Soares (2010), Andrews and Barwick (2012)). These procedures take care to ensure that the likelihood of erroneously concluding that a constraint is slack is asymptotically negligible. The researcher can then re-center the null distribution to a location consistent with the results of the pre-test, simulate the empirical distribution of the test statistic via re-sampling techniques, and finally compute the $(1 - \alpha)$ quantile of the simulated distribution to serve as the critical value for the test. Our revised statistical test now incorporates these elements because the literature shows compelling evidence that these techniques can substantially increase power without inflating asymptotic

[3] In particular, under the distribution centered at B, there is a much higher likelihood of observing data beyond the red line in comparison with the null distribution centered at A. If we take the critical value to the be the $(1 - \alpha)$ quantile of the null distribution, then the critical value that we would use under the null distribution centered at B will be larger in comparison with the one associated with the null distribution centered at A.

Figure A5: Potential Conservatism in Tests with Inequality Constraints



Notes: The constraints imposed are $\gamma_1 \geq 0$ and $\gamma_2 \geq \gamma_1$. The cone, C, shows all values in the parameter space that satisfy the inequality constraints. The plot shows two distributions of Y that are both consistent with the null hypothesis. The one centered at $(0, 0)$ is the least favorable null. The red line shows the set of points that are all the some specified distance from the cone. The dashed circles represent contour lines of the null distributions.

size. We will detail this procedure next.

The Statistical Procedure

1. We run the following unrestricted regression model:

$$y_{cj} = \beta_0 + \beta_1 X_c + \gamma_j^g + \epsilon_{cj} \tag{3}$$

where c and j denote the case and the judge, respectively. The X_c term is a vector of case characteristics including race, a set of age indicators, total counts, whether the offense is a person crime, a drug crime, violates a special rule, whether the defendant hired private counsel, plea status, the state's presumptive sentence length, presumptive sentence length-by-drug crime interaction, and a set of year fixed effects. These covariates adjust for potential imbalance in case characteristics across judges along observable dimensions.[4] The γ_j^g parameters represent a set of indicator variables for each judge-by-offender gender cell. The sum of square residuals from the unrestricted model is denoted as $\hat{\epsilon}'\hat{\epsilon}$.

2. We run a version of the regression model in (3) that imposes the following set of $\frac{k(k-1)}{2}$ inequality constraints:

$$(\gamma_j^W - \gamma_{j'}^W)(\gamma_j^M - \gamma_{j'}^M) \geq 0 \; \forall \; j \neq j' \tag{4}$$

The sum of square residuals from this restricted model is denoted as $\epsilon^{*\prime}\epsilon^*$. We can then construct a $\bar{F}$-statistic using the unrestricted and restricted sum of squared residuals

$$\bar{F} = \frac{\left(\epsilon^{*\prime}\epsilon^* - \hat{\epsilon}'\hat{\epsilon}\right)/q}{\hat{\epsilon}'\hat{\epsilon}/(n-p)}$$

where q is the number of constraints imposed in the model, and p denotes the number of parameters estimated in the regression. Note that the constrained optimization problem can be solved using fmincon, CVX, or Knitro packages in Matlab.

3. Before we simulate the distribution of the $\bar{F}$-statistic, we will conduct a series of GMS pre-tests in order to determine which of the $\frac{k(k-1)}{2}$ inequality constraints bind. In particular, we estimate the degree of slackness

[4]Even though cases are randomly assigned, there is still a benefit to conditioning on various case facts since the judicial incarceration rates will be estimated with more precision.

in the q-th constraint with the following statistic:

$$m_q \equiv \frac{1}{\kappa_n} \frac{(\hat{\gamma_j}^W - \hat{\gamma_{j'}}^W)(\hat{\gamma_j}^M - \hat{\gamma_{j'}}^M)}{\tilde{\sigma}_q}$$

where κ_n is a $o(n^{\frac{1}{2}})$ sequence of positive numbers, $\tilde{\sigma}_q$ is the bootstrapped standard error of $(\hat{\gamma_j}^W - \hat{\gamma_{j'}}^W)(\hat{\gamma_j}^M - \hat{\gamma_{j'}}^M)$, and $\hat{\gamma}_j^g$ is the unrestricted estimate of judge j's incarceration rate towards gender g felons. We compare this test statistic to 1 to determine whether the constraint binds or not.[5]

4. If we find that $m_q < 1$, then we conclude the constraint is binding, whereas $m_q > 1$ implies that the constraint is non-binding. Note that this pre-test procedure yields a different set of constraints in comparison with those in (4). We now re-run the regression model (3) but impose the inequality constraints implied by the GMS pre-test procedure. We will denote the estimates of β_0, β_1, and the judicial incarceration rates from this restricted regression as β_0^{**}, β_1^{**}, and γ_j^{g**}. In addition, we will refer to the residuals from this restricted regression as ϵ^{**}.

5. Re-sample the residuals ϵ^{**} with replacement. Construct a new outcome variable, y_{cj}^{**}, by plugging in β_0^{**}, β_1^{**}, γ_j^{g**}, and ϵ^{**} into equation (3). Note that this is analogous to re-sampling from re-centered data in a non-regression framework. For each re-sample, re-run steps 1 and 2 in order to obtain a simulated $\bar{F}$-statistic.[6] Conducting this step a large number of times will generate an empirical distribution of the $\bar{F}$-statistic from which we can compute the $(1-\alpha)$ quantile to serve as the critical value for our statistical test. We can then compare the observed $\bar{F}$-statistic to our critical value to ascertain whether rejection of the null is warranted or not.[7]

[5]The κ_n term is referred to as a tuning parameter in the literature. In our analysis, we take $\kappa_n = (lnn)^{\frac{1}{2}}$. Andrews and Barwick (2012) recommends using a *refined moment selection* procedure that computes a finite data-dependent tuning parameter. However, computation of the tuning parameters can be intensive for models with more than 10 inequality constraints. Because the analysis includes statistical tests with more than 10 constraints, we do not use this approach.

[6]The number of bootstrapped samples that we use is 1,000.

[7]It is worth emphasizing that the main interest of this empirical exercise is not the estimates of judicial incarceration rates *per se*, but what the estimates reveal about the underlying incentives that affect judicial sentencing. If we reject the null hypothesis, then this implies that judges engage in taste-based discrimination. In that sense, the spirit of our statistical test is more akin to a specification test of an economic model defined by

3 Additional Empirical Results

3.1 DiNardo, Fortin, and Lemieux (1996) (DFL) Derivation

Because many of the results in the paper rely on semi-parametric re-weighting techniques as in DiNardo et al. (1996), we present a brief overview of the methodology here. Consider an observation in our dataset represented by the vector (s, x, g), where s is the sentencing outcome, x is a vector of case characteristics, g is a gender indicator that is 1 if the felon is female and 0 otherwise. The observed joint distribution of the data is given by $f(s, x, g)$. The sentence length distribution conditional on gender can be obtained by integrating the product of the conditional distribution and the gender-specific covariate distribution over the support of x, Ω_x.

$$f(s|g) = \int_{\Omega_x} f(s|x, g) f(x|g) dx$$

We are interested in estimating counterfactual sentencing distributions; for example, the sentencing distribution that would have arisen for males if males had the same characteristics as females, $f(s_1|x, g = 0)$, where s_1 denotes the potential sentence length that a male offender would receive if the offender was female. While this counterfactual distribution is unobserved, it can be estimated by constructing a re-weighting function:

$$\begin{aligned} f(s_1|g = 0) &= \int_{\Omega_x} f(s|x, g = 0) f(x|g = 1) dx \\ &= \int_{\Omega_x} f(s|x, g = 0) \psi(x) f(x|g = 0) dx \end{aligned}$$

where $\psi(x) \equiv \frac{f(x|g=1)}{f(x|g=0)}$. Applying Bayes Rule, we can re-write the weighting function as $\psi(x) = \frac{f(g=1|x)}{f(g=0|x)} \frac{f(g=0)}{f(g=1)}$. The counterfactual density is obtained by re-weighting the covariate distribution for males by $\psi(x)$. If x is discrete, then the weighting function can be computed non-parametrically by estimating the relatively likelihood of the observation corresponding to a woman in each cell. Otherwise the weighting function can be estimated using a logit or probit. In our analysis, we run a probit model to predict the probability of the offender's gender conditional on observable characteristics. Case characteristics include

inequality constraints.

indicators for whether the defendant is represented by private counsel, plea status, person crime, total counts, whether felon violates a special rule, and indicators for black and Hispanic. We control for age by including indicators for 4 separate age groups, $< 25, 25-34, 35-44, 45+$. We also include a set of criminal severity and criminal history fixed effects and in some specifications include severity-by-age and criminal history-by-age interactions.

3.2 Females Re-weighted

Here, we present results from the semi-parametric decomposition when we re-weight females to have the same covariate distribution as males. Panel B shows these results. Panel A are the results in the paper that re-weight males to have the same distribution as females and is reproduced here for convenience. There is still a 2.5 percentage point gender gap in incarceration when we control for the full set of covariates, which translates to a 12% unexplained difference in the gender disparity. The sentence length gap falls close to zero and is not statistically significant. That the incarceration gap is smaller when we re-weight females to look like males is not surprising. This exercise will put more weight on females who share similar characteristics as male offenders. Male offenders are more likely to commit more severe crimes and have more extensive criminal histories than females. It seems reasonable that judges are less lenient towards women who commit worse crimes. In the main part of the paper, we focus on the results that re-weight males to look like females because the data is concentrated in the low-severity and low-criminal-history portion of the sentencing grid, thus the exercise has more statistical support.

3.3 Gender Difference in Judicial Case Composition

In this section, we examine the possibility that judicial heterogeneity could, in theory, be driven by gendered differences in case composition across judges rather than judicial differences in sentencing preferences. We use the aforementioned re-weighting techniques as in DiNardo et al. (1996) to address this concern. Specifically, we construct a weighting function that takes the form $\frac{P(j_0|x)}{P(j|x)}\frac{P(j)}{P(j_0)}$, where x represents a vector of case facts, j_0 is the baseline judge, and we estimate $P(j|x)$ via a probit model. Importantly, we include interactions between case facts and gender in x in order to equalize gender differences in case facts across judges. In turn, the judge effects estimated from the re-weighted model will represent heterogeneity in judicial sentencing preferences that accounts for potential gendered differences in case composition across judges.

Table A1: Semi-Parametric Decomposition of Gender Sentencing Disparities

Panel A: Males Re-weighted to Have Female Covariate Distribution

Female-Male Gap in:						
Incarceration	-0.198	-0.211	-0.163	-0.058	-0.057	-0.055
	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)
Log(Prison Months)	-0.442	-0.437	-0.175	-0.042	-0.031	-0.033
	(0.025)	(0.025)	(0.023)	(0.022)	(0.022)	(0.022)

Panel B: Females Re-weighted to Have Male Covariate Distribution

Female-Male Gap in:	Actual	(a) Age	(b) (a) + Severity	(c) (b) + Criminal History	(d) (c) + Age Intx	(e) (d) + Case Facts
Incarceration	-0.198	-0.203	-0.147	-0.006	-0.028	-0.025
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Log(Prison Months)	-0.442	-0.409	-0.156	0.031	0.008	-0.006
	(0.025)	(0.025)	(0.025)	(0.024)	(0.025)	(0.024)

Notes: These results use non-drug felony offenses only. We use probit models to predict gender probability conditional on observables. Case facts include indicators for whether the defendant is represented by private counsel, plea status, person crime, number of counts, whether felon violates a special rule, and indicators for black and Hispanic. We control for age by including indicators for 4 separate age groups, $< 25, 25-34, 35-44, 45+$.

Table A2: Judicial Heterogeneity Adjusted for Gender-by-Fact Differences

Dep Var: Incarceration

			Re-weighted Results:					
	(1) Unadjusted		(2) Case Facts		(3) (2) + Dems		(4) (3) + Facts	
Measures of Dispersion:	Male	Female	Male	Female	Male	Female	Male	Female
Standard Deviation	0.070	0.062	0.061	0.071	0.066	0.076	0.070	0.076
75/25th Percentile Difference	0.098	0.087	0.076	0.085	0.071	0.092	0.082	0.098
90/10th Percentile Difference	0.174	0.141	0.140	0.152	0.149	0.166	0.156	0.171

Note: In columns 2, 3, and 4, cases are re-weighted such that the gender difference in case composition is equalized across judges. In column 2, the case facts include presumptive sentence length, special rule violations, and person crimes. In column 3, we add demographic variables including age and race. In column 4, we add private counsel, total counts, and plea status.

Table A2 shows measures of dispersion that place more weight on judges whose judge effects are estimated more precisely. We present the standard deviation, 75/25th percentile difference, and the 90/10 percentile difference across different models and separately for men and women. The first column shows the measures of dispersion that do not adjust for case facts. These results show that there is considerable heterogeneity across judges in the raw data. A 1 standard deviation increase in judicial assignment is associated with

a 7.0 and 6.2 percentage point increase in the likelihood of incarceration for men and women, respectively. Given that the male and female incarceration rates are 0.319 and 0.121, respectively, this constitutes roughly a 22% and 51% change for men and women, respectively.

Column 2 shows measures of dispersion when we re-weight caseloads to account for gender differences in presumptive sentence length, special rule violations, and person vs. non-person offense. While some of the measures of dispersion are slightly more moderate (e.g. the 75/25 percentile difference for men falls roughly 22% from 0.098 to 0.076), their magnitudes still imply that judicial preference has substantial impact on the likelihood of incarceration. For example, movement from the 25th to 75th percentile judge is associated with a 24% increase the likelihood of incarceration for men. For women, two of three measures of dispersion (the sd and 90/10 difference) actually increase slightly. When we add demographics (age and race, column 3) and additional case facts (private counsel, total counts, and plea status, column 4) to the set of covariates in columns 3 and 4, respectively, the results are qualitatively similar.

We should note that the stability across these different models is not necessarily surprising. We have had numerous phone conversations with county clerks and judges in which we were reassured that cases are randomly assigned to judges. Indeed, in the next section, we will show that presumptive sentence length, a parsimonious proxy of case type, is balanced across judges which is consistent with random assignment. Given that cases are randomly assigned, it would be peculiar if case composition was balanced, on average, but systematically differed along gender lines. In summary, these results confirm that the observed judicial heterogeneity is robust to differences in gender-by-fact composition which is expected in light of how cases are assigned in Kansas.

3.4 Balancing Table

Table A3 shows results from balancing tests that assess whether cases are randomly assigned to judges. We regress the presumptive sentence length on judge fixed effects separately for each district and then test whether the judge fixed effects are jointly equal to zero. The first two columns uses all non-drug cases and shows the F-statistics and p-values associated with the tests. In the next two columns, we run similar tests but restrict the sample to first-time offenders. If cases are selectively assigned to judges, then there may be differentially stronger evidence of random assignment of cases involving first-time offenses to the extent that judges and prosecutors have less information on first-time offenders. However, the results are similar to those that use the full

sample of non-drug crimes. Finally, we also test whether there is any evidence of gender-based case assignment by including gender-by-judge interactions to the regressions. This will tell us if some judges are assigned cases with worse female offenders than others. The interactions are jointly significant in only 1 judicial district. It is worth noting that the latter is the balancing test that needs to be satisfied in order to validate the rank-order test.

Table A3: F-statistic associated with Joint Test of Judge Fixed Effects

Dep Var: Presumptive Sentence Length

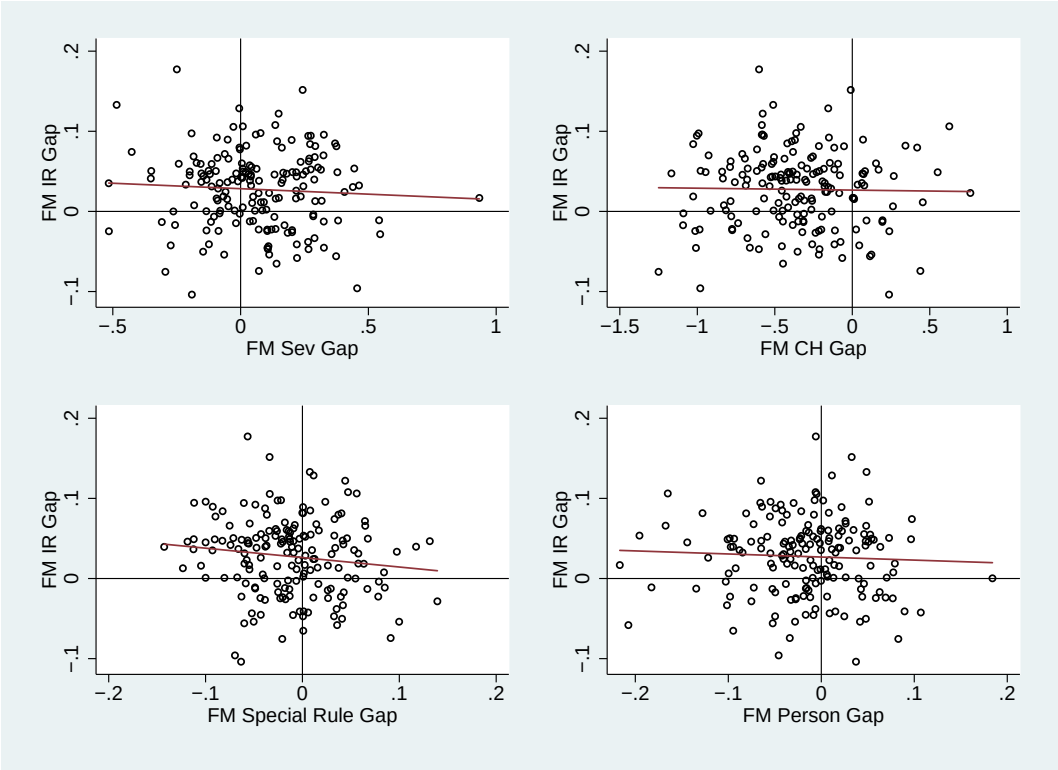
	Full Sample		First-Time Offenders		Judge-by-Gender Intx	
District	F-Statistic	P-value	F-Statistic	P-value	F-Statistic	P-value
1	8.365	0.000	2.500	0.083	0.556	0.574
2	1.841	0.138	2.738	0.043	0.015	0.997
3	1.960	0.020	2.359	0.004	0.477	0.938
4	1.903	0.149	4.287	0.014	0.290	0.748
5	4.159	0.006	0.655	0.580	0.499	0.683
6	1.957	0.119	0.344	0.794	0.063	0.979
7	0.206	0.893	1.485	0.218	0.134	0.940
8	3.275	0.006	0.553	0.736	0.436	0.823
9	0.722	0.539	0.515	0.672	0.763	0.515
10	3.648	0.000	2.226	0.014	0.603	0.813
11	2.030	0.072	2.457	0.033	0.765	0.575
12	0.314	0.575	0.220	0.640	0.116	0.734
13	0.462	0.764	1.596	0.174	0.566	0.687
14	1.262	0.278	0.210	0.958	0.163	0.976
15	1.726	0.190	0.517	0.474	2.167	0.142
16	1.778	0.149	1.390	0.246	0.742	0.527
17	.	.	.	.	.	.
18	5.673	0.000	1.958	0.002	0.664	0.914
19	1.937	0.145	1.796	0.170	0.320	0.727
20	2.055	0.104	2.341	0.073	0.426	0.734
21	0.589	0.555	0.351	0.704	0.010	0.990
22	2.688	0.102	0.077	0.782	0.113	0.736
23	1.456	0.234	2.441	0.089	0.443	0.642
24	3.755	0.053	0.729	0.395	1.091	0.297
25	2.412	0.065	1.507	0.212	0.268	0.849
26	2.741	0.027	1.806	0.127	2.323	0.055
27	15.601	0.000	2.555	0.038	2.818	0.024
28	1.501	0.186	0.890	0.487	0.401	0.848
29	9.608	0.000	2.836	0.000	0.883	0.577
30	1.221	0.295	3.974	0.020	0.789	0.455
31	0.966	0.425	0.998	0.408	0.257	0.906
# of Districts (p-value $\leq$ 0.05)	9		9		1	

Notes: These results use non-drug felony offenses only. We run regressions of presumptive sentence length on judge fixed effects separately by district. We test whether the judge fixed effects are jointly equal to zero. F-statistics and associated p-values are reported. The second column restricts the sample to first time felons and in last column, we include judge-by-gender interactions are report F-statistics and p-values associated with the test that the interactions are jointly equal to zero.

We also present graphical evidence that cases are not assigned to judges by gender. Figure A6 plots the relationship between the judge-specific female-

male incarceration gap and the judge-specific female-male severity, criminal history, special rule, or person crime gap in the cases appearing before each judge.

Figure A6: Judge-Specific FM Disparity in Incarceration and Case Facts



Notes: Each dot represents judge-specific gender disparity relative to the baseline judge, who is normalized to 0. The horizontal and vertical lines pass through the baseline judge. The estimates 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.

While there is evidence of heterogeneity in the female-male incarceration gap, there is little evidence this is driven by female-male differences in case assignment. Consider the figure in the upper left. If it were the case that judges who punish women more harshly (and thus the female-male gap is more positive) happen to be assigned worse female offenders (with worse severity measures), then we would expect an upward slope to these figures. Instead, all of the estimated relationships are flat and none are statistically significant from zero at the 5% level. This corroborates that the heterogeneity in judicial treatment towards females is unlikely to be driven by gender differences in the

distributions of severity, criminal history, special rules violations, and person crimes across judges.

3.5 Judicial Entry and Changes in Case Composition

In this section, we examine the possibility that our event-study analysis may reflect changes in case composition due to the entry of a harsh or lenient judge. There are numerous reasons why the entry of a harsh or lenient judge could lead to changes in case composition. For example, the presence of a harsh judge may lead prosecutors to file charges for marginal cases, deter criminals from re-offending upon release, or encourage police to pursue certain arrests more aggressively when the expected punishment is higher. These types of behavioral responses could lead to changes in the severity of the crimes, the criminal history of the offenders, and the types of charges or special rules violations that are applied, which in turn could affect sentencing.

However, it is less clear as to how these hypothetical changes should affect the interpretation of our results. If the changes in case composition are a direct response to the entry of the harsh or lenient judge, then the increase in incarceration rates are ultimately driven by changes in judicial composition. In this case, the change in case composition would point to a mechanism rather than a confound for our analysis. Nonetheless, our view is it would be a worthwhile exercise to examine the degree to which case facts change with respect to judicial composition. To this end, we will provide a new set of results from the following regression model:

$$y_{idt} = \gamma_t + \tau_d + \delta Post_{idt} + X_i\beta + \epsilon_{idt} \tag{5}$$

where y_{idt} is the outcome variable, γ_t and τ_d are a set of year and district fixed effects, respectively, X is the usual set of case facts, and $Post_{idt}$ is a binary variable that indicates whether the case is sentenced before or after the entry or exit of a lenient or harsh judge. Harsh and lenient judges are defined in terms of their incarceration rate of men who commit non-person offenses. This regression is restricted to women only. The subscripts i denote the case, d denotes the district, and t reflects the year. Districts that do not experience an event of entry or exit will identify the year fixed effects. Standard errors are clustered at the district-level.

There are two important differences between this regression model and the event-study model employed in the paper. First, the outcome variable is no longer incarceration. Instead, our primary specification of interest will use a measure of **predicted incarceration** ($\widehat{incar}$) as the outcome variable.

Predicted incarceration represents the fitted values from a regression of incarceration on the usual set of case facts, the criminal severity level, criminal history level, race, age, plea status, private counsel, and total counts. We will then regress $\widehat{incar}$ on $Post_{idt}$ and γ_t and τ_d. In addition, we will show results from models that separately examine how various case facts change pre vs. post entry or exit of a harsh or lenient judge. These results will provide more context for the observed changes in $\widehat{incar}$.

Second, notice that the key independent variable is now $Post_{idt}$ which collapses the set of timing indicators into a single variable. While this specification is less flexible in that it does not allow us to observe the time pattern in case facts, it is a much more parsimonious way of presenting the results which we prefer due to the large number of case facts that we examine.

Table A4: Case Mix and Composition of Judiciary

Coefficient on Post Entry or Exit

	"Harsh" Judge		"Lenient" Judge	
Outcome Variables:	Entry	Exit	Entry	Exit
Incarceration	0.074***	0.014	-0.058***	-0.020
	(0.010)	(0.014)	(0.013)	(0.013)
Predicted Incarceration	-0.043***	-0.039***	0.018	0.014*
	(0.009)	(0.013)	(0.012)	(0.007)
Case Facts:				
Log(Presumptive Sentence Length)	-0.179***	0.040	-0.009	0.004
	(0.026)	(0.059)	(0.027)	(0.024)
Severity	-0.366***	0.112	-0.030	0.015
	(0.050)	(0.078)	(0.058)	(0.050)
Criminal History	-0.064	-0.306	0.188**	0.089
	(0.075)	(0.272)	(0.075)	(0.089)
Special Rule Violation	0.023	-0.101*	0.058**	0.046**
	(0.014)	(0.052)	(0.026)	(0.020)
Person Crime	0.021*	0.027*	-0.022*	-0.015*
	(0.012)	(0.014)	(0.012)	(0.008)
Log(Total Counts)	0.083***	-0.044	-0.020	-0.016
	(0.015)	(0.039)	(0.022)	(0.015)
Private Counsel	0.049**	0.015	-0.006	0.015
	(0.020)	(0.107)	(0.017)	(0.045)
Plea	-0.032***	-0.002	0.007	0.005
	(0.004)	(0.013)	(0.008)	(0.008)

Notes: Predicted incarceration represents the fitted values from a regression of incarceration on the usual set of case facts, the criminal severity level, criminal history level, race, age, plea status, private counsel, and total counts.

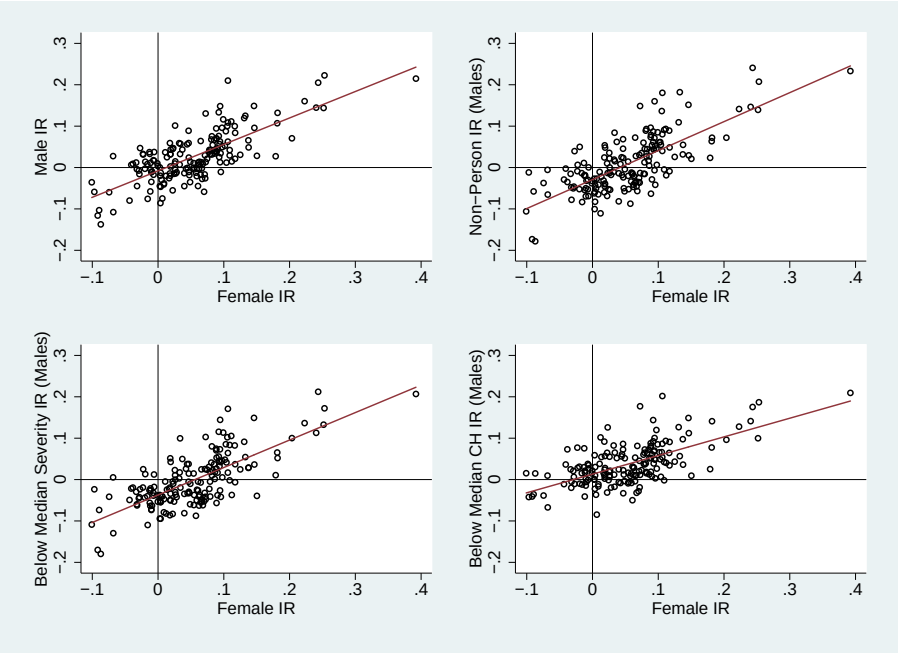
Table A4 shows the results. The first row shows results that are qualitatively similar to those in the previous version of the paper. The incarceration rate of women increases (decreases) when a harsh (lenient) judge enters, but there is relatively little impact when a harsh or lenient judge exits. The key set of results are in the second row. In column 1, the estimate implies that the entry of a harsh judge is associated with a subsequent *decrease* in predicted incarceration. This suggests that the 7.4 percentage point increase in female incarceration coincides with a time period when women are relatively less felonious. The analysis of case facts shows that the decrease in predicted incarceration is driven by less severe offenses and an increase in private counsel. Similarly, in column 3, there is little evidence that the 5.8 percentage point decrease in female incarceration is driven by case facts. In fact, the coefficient on predicted incarceration is positive but not statistically significant. While it is possible that case mix could, in theory, respond to judicial composition, this analysis does not lend strong support to this hypothesis.

3.6 Judicial Heterogeneity in Sentencing Preferences

Our prior is that judges who are "tough on females" might also be "tough on crime" in general. To assess this, Figure A7 presents a series of plots that show the correlation between the judge-specific female incarceration rate and other judge-specific incarceration rate. Moving left-to-right, the other judge-specific incarceration rates in the top row include male incarceration rates and male incarceration rates among non-person crimes, and in the bottom row, male incarceration rates among below-median severity crimes, and male incarceration rates among below-median criminal history offenders. These "other" judicial incarceration rates focus on male offenders to avoid conflating severity effects or criminal history effects with gender. Recall that these lower severity crimes and criminal histories are consistent with the records of female offenders. Each panel shows how one of the other judicial incarceration rates correlates with judicial female incarceration rates. All of the incarceration rates are regression adjusted for the usual set of covariates.

All four panels of Figure A7 exhibit a strong positive correlation. Judges who incarcerate female offenders at high rates also tend to have high incarceration rates of males, and of males who have sparse criminal histories, males who commit non-violent and less severe crimes. Bivariate regressions confirm that the estimated relationships are highly statistically significant at conventional levels. The figure also reinforces that there is considerable judicial heterogeneity in both male and female incarceration rates, as there is comparable variation along both the vertical and horizontal axis.

Figure A7: Judicial Heterogeneity by Gender and Type of Crime



Notes: 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.

3.7 Prison Diversion

It is worth noting that Kansas did implement a prison diversion program for drug offenders during our sampling period. Our sense is that the prison diversion program will not affect our main results. This is because most of our empirical work - quantifying judicial heterogeneity, the event-study analysis, and the rank-order test - focuses exclusively on non-drug related offenses. In Kansas, the overwhelming majority of cases that are eligible for the prison diversion program (e.g. close to 99%) are drug-related offenses. Thus, our main analysis does not use cases directly affected by this program. However, it is possible that our analysis is affected to the extent that case assignment to judges responded to the program. Through conversations with district court clerks and judges, we have learned that the introduction of the prison diversion program did not have an impact on how cases were assigned to judges. Once a case enters the system, it is assigned to a judge via a computer algorithm regardless of whether or not it is a non-drug or drug related offense. This is

consistent with our finding that the presumptive sentence length is balanced across judges in the majority of judicial districts in Kansas.

Table A5: Gender Disparity Excluding Drug Court Cases

Restricted to Drug Related Offenses				
Female-Male Gap in:	(1)	(2)	(3)	(4)
Incarceration Rates	-0.152***	-0.161***	-0.066***	-0.070***
	(0.007)	(0.008)	(0.008)	(0.008)
Log(Prison Length)	-0.129***	-0.173***	-0.089***	-0.082***
	(0.030)	(0.029)	(0.023)	(0.022)
Covariates:				
Race	N	Y	Y	Y
Age	N	Y	Y	Y
Severity Level	N	N	Y	Y
Criminal History	N	N	Y	Y
Case Facts	N	N	N	Y

Note: There are 29,698 drug-related cases in which the offender was not sentenced to a prison diversion programs. Of these, 7,896 cases resulted in a prison term.

Nonetheless, it seems prudent to explore the degree to which the prison diversion program may have affected our results. The analysis that is most likely to be influenced by the program is our estimation of the gender sentencing disparity in drug related cases. Recall that this is the only part of empirical work that focuses on drug-related offenses and it is true that women are more likely to be assigned to a prison diversion program conditional on the usual set of case facts. Table A5 shows the estimated gender disparity in incarceration and log of the prison term among drug-related cases, but unlike the previous draft of the paper, these models restrict the sample to cases in which the offender is not sentenced to the prison diversion program. The four columns show estimates from specifications that vary what case facts are in the conditioning set. The results show striking similarity to those in Table 4 of the paper. These estimates reinforce the notion that our empirical analysis is unlikely to be biased by the introduction of the prison diversion program.

References

Andrews, D. W. and P. J. Barwick (2012). Inference for parameters defined by moment inequalities: A recommended moment selection procedure. *Econometrica 80*(6), 2805–2826.

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.

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.

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

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.

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

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