

SIMON FRASER UNIVERSITY

Department of Economics

Discussion Papers

07-12

**Statistical Discrimination
in the Criminal Justice
System: The Case for
Fines Instead of Jail**

Phil Curry and Tilman Klumpp

June 2007



Statistical Discrimination in the Criminal Justice System: The Case for Fines Instead of Jail*

Philip A. Curry[†]
Simon Fraser University

Tilman Klumpp[‡]
Emory University

June 3, 2007

Abstract

We develop a model of statistical discrimination in criminal trials. Agents carry publicly observable labels of no economic significance (race, etc.) and choose to commit crimes if their privately observed utility from doing so is high enough. A crime generates noisy evidence, and defendants are convicted when the realized amount of evidence is sufficiently strong. Convicted offenders are penalized either by incarceration or by monetary fines. In the case of prison sentences, discriminatory equilibria can exist in which members of one group face a prior prejudice in trials and are convicted with less evidence than members of the other group. Such discriminatory equilibria cannot exist with monetary fines instead of prison sentences. Our findings have implications for potential reforms of the American criminal justice system.

Keywords: Statistical discrimination, criminal justice, stereotypes, prejudice, double standards.

JEL code: D72, D78.

*We thank David Bjerk, Amie Broder, Hugo Mialon, Paul Rubin, David Scoones, Xuejuan Su, as well as participants at the 2006 Meetings of the Canadian Law and Economics Association, and the 2007 Meetings of the American Law and Economics Association, for fruitful discussions.

[†]Department of Economics, Simon Fraser University. 8888 University Dr., Burnaby, BC, V6A 1S6, Canada. E-mail: pcurry@sfu.ca.

[‡]Department of Economics, Emory University. Rich Building 316, 1602 Fishburne Dr., Atlanta, GA 30322, USA. E-mail: tklumpp@emory.edu.

“It is in justice that the ordering of society is centered.”

— Aristotle

1 Introduction

Criminal conviction rates in the United States differ drastically across racial groups. The jail incarceration rates for U.S. blacks, for example, is 800 people per 100,000, while the rate for whites is 166 per 100,000. An estimated 12% of black males in their late twenties were incarcerated in 2005, as opposed to 1.7% of white males.¹ That these differences exist is undisputed. It is much less clear, however, *why* they exist. One possible explanation is that blacks commit more crimes than whites. For example, criminal participation tends to be correlated with economic characteristics such as income, education, or area of residence, and these characteristics differ across racial groups. If this were the whole story, then empirical studies of the determinants of crime should find race to be insignificant once these other characteristics are controlled for, and this is generally not the case.² Another possibility is that the criminal justice system is somehow “biased,” so that blacks are more easily convicted. This explanation, however, begs some other questions. How could such a bias persist? And, what policy implications are there?

Some papers have examined the issue of bias in the courts. Georgakopoulos (2004) and Burke (2007) both consider the possibility that courts may have *false* beliefs about a group’s proclivity for criminality. Georgakopoulos notes that such beliefs can lead to greater arrest and conviction rates, which might reinforce such beliefs while Burke considers the psychological basis for false beliefs in prosecutors and proposes practices and institutions to prevent such cognitive bias from arising. Other papers have considered racial profiling by the police. Examples include Knowles, Persico and Todd (2001), Persico (2002), Alexeev and Litzel (2004) and Bjerk (2007). These papers, however, assume ex-ante differences in criminal behavior across racial subgroups; they are not concerned with how such difference might come to be.

In this paper, we present a theoretical model of crime in which any and all differences across groups, as well as judicial biases, arise as equilibrium results. We consider an environment in which a judge or jury must determine whether a given amount of evidence is sufficient to convict a defendant. This judge starts with prior beliefs about the defendant’s guilt, which depends on observable characteristics such as income or race, and uses the evidence to update these beliefs according to Bayes’ Rule. In equilibrium, we impose that the judge’s prior beliefs are correct in that the probability that the judge attaches to the defendant being guilty is exactly equal to the proportion of the defendant’s type that commit crime. Suppose now that convicted offenders are

¹Source: Bureau of Justice Statistics Prison and Jail Inmates at Midyear 2005.

²See, for example, Bjerk (2006), Krivo and Peterson (1996), Raphael and Winter-Ebmer (2001), and Trumbull (1989).

sent to prison for a fixed amount of time. Such a sentence punishes richer individuals more severely than poorer ones, so that poorer individuals will be less deterred by this punishment and hence be more willing to commit crimes. If, at the same time, a person's perceived proclivity to crime negatively affects their earning opportunities, one can see how a stereotyping equilibrium might arise in which members of different racial groups receive different (but correct) prior beliefs regarding their guilt. Thus, even with the same amount of evidence, the disadvantaged group will be more easily convicted. Such a theory of discrimination is often associated with the term *statistical discrimination*.

Statistical discrimination in the context of labor markets has been studied, for example, in Phelps (1972), Arrow (1973), and Coate and Loury (1993).³ Our model is different from these in that we incorporate the role of the courts. Interestingly, the policy implications of such a model differ significantly from the ones derived in labor-market models. Coate and Loury (1993), for example, propose that negative stereotyping equilibria can be prevented through affirmative action in the labor market. In our framework, stereotyping equilibria can arise even if observed wages and employment rates are the same for all individuals *not in jail* (this is done in a dynamic extension of the model). Thus affirmative action policies targeting the labor market may not have the desired effect. We identify an alternative remedy, namely the adoption of monetary fines instead of incarceration as a means of punishment. Fines tend to punish poorer persons more severely than richer ones, and hence act in the opposite way as prison sentences. In particular, we show that they eliminate stereotyping equilibria. Thus, our paper contributes to the literature on optimal sanctions, specifically on the use of fines versus incarceration.⁴ We remark that in this paper we do not consider the costs to either crime or corrections, and compare prison sentences and fines strictly with respect to the discrimination question. Hence, our paper does not offer a welfare analysis, and neither is this our intent. There are many reasons, besides the potential for biased outcomes, why one form of punishment may be preferable over another. For instance, we abstract from the idea that a defendant might be “debt proof” (unable to pay the fine), a very practical reason that incarceration might be preferred. However, we argue that, while fines may not be desirable (or even possible) for *some* crimes, this is not the case for *all* crimes.

Related to our approach is a paper by Verdier and Zenou (2004), who examine a model of statistical discrimination, location choice, and criminal activity. As in our model, race serves as a coordination device in their paper, assigning different rational expectations equilibria to different racial groups. However, their model is much different

³The idea that economically meaningless events, such as racial labels, can coordinate individual actions and social expectations is often referred to as a “sunspots theory”. Various related notions have appeared in Cass and Shell (1983), Aumann (1987), Forges (1986), and Cartwright and Wooders (2006), among others.

⁴See, for example, Becker (1968), Polinsky and Shavell (1984), Morris and Tonry (1990), Posner (1992) and Levitt (1997).

from ours in the nature of the equilibrium feedbacks between beliefs, criminal activity, and economic variables. Discriminatory outcomes arise in Verdier and Zenou because lower wages induce residential location choices closer to high-crime areas, and therefore lower the cost of committing a crime. In our model, on the other hand, it is the fact that imposed sentences affect persons with different incomes differently, which provides the feedback from economic variables to criminal activity. Furthermore, jury bias is an essential ingredient in our model that is absent in Verdier and Zenou: Here, not only do racial groups differ in their crime rates, but *conditional on having committed a crime* a member of the disadvantaged group is more easily convicted. Finally, unlike our paper, it is not possible in Verdier and Zenou’s model to generate a situation in which all working persons have the same wage rate, and nevertheless there are ex-post differences in crime rates across races.

We proceed as follows. Our theoretical model is developed in three steps. In Section 2, we assume (temporarily) that incomes are exogenously given and observable. In Section 3, we derive a number of results for this case. In Section 4, we do away with the assumption that income is exogenous—instead, we introduce different (racial) groups which are ex-ante identical, and make each group’s income a function of prejudice toward that group. We show that this can create discriminatory equilibria if prison sentences are used as punishments, but not when fines are used. In Section 5, we consider two extensions of the model. First, by making the model dynamic we demonstrate that group membership can predict crime even when per-period income can not. As mentioned above, this result casts some doubt on the success of labor market interventions to prevent statistical discrimination. Second, we examine the case where an individual may be convicted of multiple offenses, and show that the intuition from the previous sections still hold even without interaction with the labor market. Section 6 concludes with a few remarks. Most proofs are in the Appendix.

2 A Simple Model of Crime and Prejudice

In our model, an agent must decide whether to commit a crime, and a judge or jury must decide whether to convict a defendant accused of committing a crime. We assume that the agent is characterized by an exogenously given, publicly observable type $w \in [\underline{w}, \infty)$, with $\underline{w} > 0$. The type w represents an agent’s wealth or income and may affect the jury’s beliefs.⁵

⁵The exogeneity of w is a temporary assumption—we will do away with it in due time and replace it with fixed and publicly observable race labels (that are not connected to payoffs except through the behavior of agents). However, because race affects prejudice through economic variables, we defer the introduction of race until Section 4, and focus on economic variables in this section and the next.

2.1 Timing of events

The timing is as follows: First, the agent comes across an opportunity to commit a crime. The benefit to committing this crime is $\eta \in [0, \infty)$, which is privately observable and drawn according to a continuous distribution Q with support $[0, \infty)$. This distribution is assumed to be independent of w . After the agent observes η , he decides whether to commit a crime ($d = 1$) or not ($d = 0$). If the crime is committed, the agent receives the benefit of the crime, η , and an investigation is initiated. If the agent does not commit the crime, there is a probability $\lambda \in (0, 1)$ that an investigation is initiated “by accident.” If this happens, the agent does not consume η but may still be found guilty of the crime.

If under investigation, a random amount of evidence $t \in [0, 1]$ against the agent will be discovered. In case a crime has in fact been committed, t is a random draw from distribution F . In case of accidental investigation, t is drawn from distribution G . We assume that F and G have support $[0, 1]$, are continuous with density f and g , respectively, and that the ratio $f(t)/g(t)$ increases strictly in t . A higher value of t hence means stronger evidence against the agent. We further make the technical assumptions that $0 < f(0) < \infty$ and $0 < g(0) < \infty$.

After the evidence is discovered, the agent has to stand trial and becomes a defendant. At trial, a judge observes w as well as t and forms belief $\theta(w, t) = P[d = 1|w, t]$, representing the probability of guilt of the agent. The agent is convicted of the crime if $\theta(w, t) \geq \alpha$, where $\alpha \in (0, 1)$ represents the standard of proof. The interpretation of α is that courts must determine whether or not the evidence establishes the defendant’s guilt beyond a “reasonable doubt.” In our model, α quantifies what is “reasonable.” If the agent is not investigated, or if he is investigated and subsequently acquitted, he receives utility $u(w)$. Regarding the utility function u , we assume it is twice differentiable with $u'(w) > 0$ and $u''(w) \leq 0$. If the agent is convicted, he is sentenced to a punishment which reduces his utility by $\Delta(w)$ on him. We will describe the possible penalties available in more detail below. Figure 1 depicts the timing of events in a “game tree” (the agent’s payoffs are given at the terminal nodes of the tree).

It is worth discussing our assumption about accidental investigations at this point. From a technical perspective, it introduces the possibility that a person who faces trial is innocent. If this possibility did not exist, the jury’s prior beliefs and the updating problem would become trivial (every defendant would be guilty). From a conceptual perspective, there are several interpretations. One is that the agent is charged with a crime that has in fact happened, but which was committed by a third person. This interpretation would introduce some complications to our model. For instance, it would then be possible that one person can be convicted of two crimes, one that he committed and one that he did not commit. This interpretation will be examined in Section 5.2. Here, we consider a simpler model instead, in order to make the intuition for our results clearer. The current model entails an agent who must decide whether to commit an act knowing that he may be held responsible even if he did not commit it. A scenario where

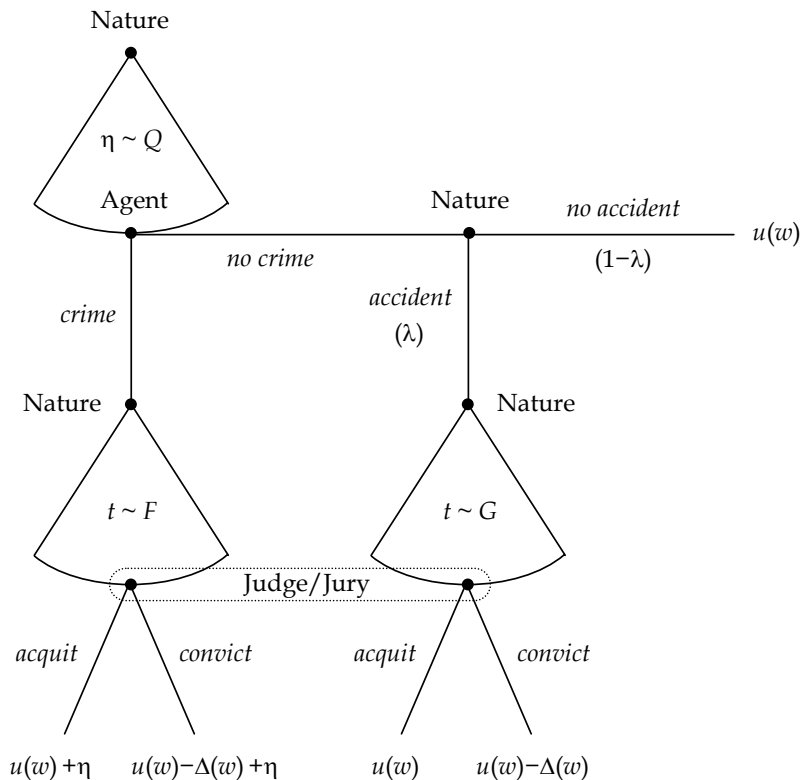


Figure 1: Timing of events

this applies to is tort law. For instance, damaging events may take place which may be the result of negligence, or not, and at trial it must be determined whether the agent should be held responsible. In the case of accidental torts, η can be interpreted as the utility benefit from *not* taking precautions; for intentional torts, η may represent some other benefit. Consider, for example, the case of a corporate bankruptcy which causes losses to the shareholders. The bankruptcy can be the result of fraudulent actions on part of the management, a crime that might have lead to greater income for those involved, but it can also simply be the result of bad business decisions or unforeseen shocks to the economic environment, neither of which would lead to liability on the part of the management. The probability λ in the model would then represent the probability of such unforeseen shocks and the subsequent investigation of the company's management.

2.2 Penalties

We now consider the different forms of punishments more explicitly. A *penalty schedule* is a mapping $\rho : [w, \infty) \rightarrow [w, \infty)$ such that $\rho(w) \leq w$. When sentenced to a penalty ρ , the agent's wealth will be decreased by $\rho(w)$. The utility loss to the agent is therefore

$\Delta(w) = u(w) - u(w - \rho(\delta))$. Listed below are a few important cases of penalties.

Imprisonment. If sentenced to prison, the agent's wealth will be reduced to a uniformly low level w_0 , which is independent of the agent's type outside of prison. Thus, a prison sentence is given by the penalty schedule $\rho(w) = w - w_0$. The utility loss to the agent is therefore $\Delta(w) = u(w) - u(w_0)$.

Simple fine. If sentenced to pay a simple fine, the agent's wealth will be decreased by a fixed amount δ . This is described by the penalty schedule $\rho(w) = \delta$, and the utility loss to the agent is $\Delta(w) = u(w) - u(w - \delta)$.

Proportional fine. If sentenced to pay a proportional fine, the agent's wealth is reduced by a fraction $\gamma \in (0, 1)$; thus $\rho(w) = \gamma w$. The utility loss to the agent is therefore $\Delta(w) = u(w) - u((1 - \gamma)w)$.

Since the utility loss Δ depends on the type of punishment being used, different penalty forms have different effects on individuals of lower and higher income types. In general, it depends on both the properties of the utility function u as well as on properties of the penalty schedule ρ whether higher or lower types are penalized more severely. However, if punishment is imprisonment, Δ increases in w , so a prison sentence clearly is a more severe punishment for higher types than it is for lower types. If the punishment is a fine δ , on the other hand, Δ decreases in w as long as agents are strictly risk averse. In this case, a simple fine works exactly in the opposite direction as a prison sentence.

2.3 Bayesian beliefs at trial

At trial, the judge forms the posterior belief $\theta(w, t)$ through Bayesian updating. Specifically, let $p(w)$ denote the judge's prior belief that an individual of type w commits a crime⁶. We will call $p(w)$ the *prejudice* held against individuals of type w . The Bayesian likelihood that the investigated individual of type w is guilty, conditioning on evidence t , is then

$$\theta(w, t) \equiv P[d = 1|w, t] = \frac{p(w)f(t)}{p(w)f(t) + \lambda(1 - p(w))g(t)} \in [0, 1]. \quad (1)$$

Given that $f(t)/g(t)$ increases, $\theta(w, t)$ increases in t and in $p(w)$. For a fixed α , let $t(w)$ be such that

$$\alpha = \theta(w, t(w));$$

$t(w)$ is then the *conviction threshold* that is applied to individuals of type w . While α is fixed for all defendants, the amount of evidence $t(w)$ required to prove guilt beyond probability α can depend on income w .

⁶Note that this prior is not subjective. In equilibrium, it is equal to the probability that an individual of type w actually commits a crime. As such, judges will not differ in their priors.

2.4 The defendant's decision

Let $m_1(w)$ denote the probability that an individual of type w is convicted conditional on having committed the crime, and let $m_0(w)$ denote the probability that the same individual is (wrongfully) convicted conditional on not having committed the act. The crime is committed if and only if the benefit from doing so exceeds its cost:

$$\eta > q(w) \equiv [m_1(w) - m_0(w)] \Delta(w),$$

where $\Delta(w)$ is the utility loss of the agent when punished, and the difference $m_1(w) - m_0(w)$ is the increase in the likelihood of suffering this loss when committing the crime. The product $q(w) = [m_1(w) - m_0(w)]\Delta(w)$ is then the expected cost of committing the crime, so that the defendant decides to commit the act when the benefit of doing so, η , exceeds the expected cost, $q(w)$. We can express the probability of conviction following a crime from the agent's perspective as

$$m_1(w) = P[\theta(w, t) \geq \alpha | d = 1] = P[t \geq t(w) | d = 1] = 1 - F(t(w)),$$

and the probability of wrongful conviction as

$$m_0(w) = \lambda P[\theta(w, t) \geq \alpha | d = 0] = \lambda P[t \geq t(w) | d = 0] = \lambda(1 - G(t(w))).$$

2.5 Rational expectations equilibrium

An equilibrium will be a tuple (p, q, t) , where $p : [\underline{w}, \infty) \rightarrow [0, 1]$ is the prior belief for the judge, $q : [\underline{w}, \infty) \rightarrow [0, \infty)$ is the decision threshold for the agent, and $t : [\underline{w}, \infty) \rightarrow [0, 1]$ is the conviction threshold; all of these are functions of the agent's type. We call (p^*, q^*, t^*) an *equilibrium* if it solves the following system of equations for all w :

$$p^*(w) = 1 - Q(q^*(w)), \tag{2}$$

$$q^*(w) = [1 - F(t^*(w)) - \lambda(1 - G(t^*(w)))] \Delta(w), \tag{3}$$

$$\theta(w, t^*(w)) = \alpha. \tag{4}$$

Condition (2) says that in equilibrium, the prejudice toward a defendant of type w , $p^*(w)$, is consistent with the probability that agents of type w actually commit crimes. Condition (3) says that the decision of a type w agent to commit a crime, given by $q^*(w)$, is optimal given the conviction thresholds $t^*(w)$ applied to this agent. Condition (4) says that the conviction threshold be such that a conviction occurs if and only if the evidence establishes the defendant's guilt beyond probability α , where this probability is computed by Bayes' Rule using the judge's prejudice $p^*(w)$ as the prior. The equilibrium is hence one of *rational expectations*. We begin with establishing existence.

Lemma 1. *A rational expectations equilibrium exists.*

Note that Lemma 1 does not preclude the existence of multiple equilibria. If multiple equilibria exist, then it is obvious that groups that do not differ in their economic fundamentals could differ in their criminal behavior: One group could simply be in a low crime equilibrium while the other is in a high crime equilibrium. In order to explain why groups differ in their equilibria, one could then appeal to a story as in Sah (1991). If one group had historical reasons to have higher crime rates, say because they were historically poorer, then differences in crime rates could persist even after differences in the economic fundamentals were eliminated.

Of more interest, however, is that such differences can arise even when there exists a unique equilibrium outcome to the model as thus far described. (This, of course, requires an additional feedback channel, namely from beliefs to income, which we will introduce in Section 4.) We therefore continue by finding a condition for uniqueness. In order to do so, we define

$$\bar{\lambda} \equiv f(0)/g(0).$$

Given our assumptions on f and g , it will be the case that $0 < \bar{\lambda} < 1$. We then have the following result:

Lemma 2. *If $\lambda \leq \bar{\lambda}$, the rational expectations equilibrium is unique.*

Below, we will be concerned with how the equilibrium values for $p^*(w)$, $q^*(w)$, and $t^*(w)$ vary with w . Note that w enters the equilibrium definition directly only in condition (3), i.e. the condition stating that the agent's crime decision be optimal. Suppressing the dependence on w , this condition is

$$q = [1 - F(t) - \lambda(1 - G(t))]\Delta.$$

Intuitively, we expect q to increase in Δ and decrease in t . Note that $1 - F(t) - \lambda(1 - G(t))$ is weakly decreasing in t if and only if $\lambda \leq f(t)/g(t)$. Since $f(t)/g(t)$ is increasing by assumption, if $\lambda \leq \bar{\lambda}$ then $1 - F(t) - \lambda(1 - G(t))$ is non-increasing for all t . For all $\lambda \leq \bar{\lambda}$, we therefore get

$$\lambda \leq \bar{\lambda} \Rightarrow \frac{\partial}{\partial t} q = [-f(t) + \lambda g(t)] \Delta \leq 0. \quad (5)$$

Thus, if $\lambda \leq \bar{\lambda}$, an increase in the conviction threshold t indeed leads to a decrease in the decision threshold q of the agent. Furthermore, since $1 - F(1) - \lambda(1 - G(1)) = 0$, we have

$$\lambda \leq \bar{\lambda} \Rightarrow \frac{\partial}{\partial \Delta} q = 1 - F(t) - \lambda(1 - G(t)) = - \int_t^1 [-f(s) + \lambda g(s)] ds \geq 0. \quad (6)$$

Thus, if $\lambda \leq \bar{\lambda}$, an increase in the potential penalty Δ leads to an increase in the decision threshold q of the agent.

These preliminary observations allow us to characterize the equilibrium further. In particular, we will state several results concerning prejudice. We call an equilibrium *biased against lower types* if

$$w > w' \Rightarrow p^*(w) < p^*(w'), t^*(w) > t^*(w'), \text{ and } q^*(w) > q^*(w').$$

That is, the judge is prejudiced in favor of higher types and applies a higher conviction threshold to higher types, and higher types are less likely to commit a crime. Similarly, an equilibrium *biased against higher types* if the reverse inequalities hold:

$$w > w' \Rightarrow p^*(w) > p^*(w'), t^*(w) < t^*(w'), \text{ and } q^*(w) < q^*(w').$$

Finally, if p^* , q^* and t^* are constant, the equilibrium is *unbiased*. hold:

$$w > w' \Rightarrow p^*(w) > p^*(w'), t^*(w) < t^*(w'), \text{ and } q^*(w) < q^*(w').$$

Finally, if p^* , q^* and t^* are constant, the equilibrium is *unbiased*.

3 Penalties and Equilibrium Bias

We begin with a general result concerning the potential bias that can arise in equilibrium. (As before, we assume that the defendant's income w is exogenously given and observed by the jury.)

Lemma 3. *Suppose $0 < \lambda \leq \bar{\lambda}$, and let (p^*, q^*, t^*) be the unique equilibrium. The equilibrium is (i) biased against higher types when Δ strictly decreases in w , (ii) biased against lower types when Δ strictly increases in w , and is (iii) unbiased if Δ is constant.*

With this result in mind, we are now ready to compare different forms of punishment with respect to whether they lead to biased equilibria or not, and if they do, whether the bias is against lower types or against higher types. In our analysis, punishments differ only in terms the utility loss $\Delta(w)$, they impose on defendants of type w . The following results are therefore all direct consequences of Lemma 3.

We begin by describing the general penalty schedule that leads to unbiased equilibria. By Lemma 3, the equilibrium is unbiased if Δ is a constant, i.e. $\Delta(w) = \bar{\Delta}$ or $\Delta'(w) = 0 \forall w$:

$$\Delta'(w) = u'(w) - (1 - \rho'(w))u'(w - \rho(w)) = 0, \tag{7}$$

so that unbiasedness requires ρ to satisfy the following differential equation and initial condition:

$$\rho'(w) = 1 - \frac{u'(w)}{u'(w - \rho(w))}, \tag{8}$$

$$\rho(\underline{w}) = \underline{w} - u^{-1}(u(\underline{w}) - \bar{\Delta}), \tag{9}$$

Condition (9) describes the punishment applicable to type \underline{w} to achieve the desired utility loss $\bar{\Delta}$. The differential equation (9) then describes how to trace out the penalty schedule ρ that maintains the same utility loss for all types.⁷ This particular penalty schedule, given in (9), describes a knife-edge case in that it characterizes those penalties which lead to unbiased equilibria. Before turning to biased equilibria, we show that the knife-edge case can be achieved by simple and proportional fines, respectively, if the utility function u satisfies particular properties:

Theorem 4. *The equilibrium is unbiased in the following special cases:*

- (a) *Punishment is by a simple fine and agents are risk-neutral, i.e. $u''(w) < 0 \forall w$.*
- (b) *Punishment is by a proportional fine and agents have constant relative risk aversion of 1. That is, $\rho(w) = \gamma w$ for some $\gamma \in (0, 1)$, and $u(w) = a \ln w + b$ for some b and some $a > 0$.*

Let us now consider under which conditions the equilibrium is biased. By Lemma 3, $\Delta'(w) > 0$ implies that the equilibrium will be biased against lower types. We can therefore derive a condition similar to (9), that is, the equilibrium is biased against lower types if the penalty schedule satisfies

$$\rho'(w) > 1 - \frac{u'(w)}{u'(w - \rho(w))}.$$

Likewise, it is biased against higher types if the reverse inequality holds,

$$\rho'(w) < 1 - \frac{u'(w)}{u'(w - \rho(w))}.$$

Of course, given any schedule ρ , it need not be the case that the equilibrium is monotone in w , as the resulting $\Delta(w)$ may not be monotone in w . As far as prison sentences and simple fines are concerned, however, we can make the following statement:

Theorem 5. *Suppose $0 < \lambda \leq \bar{\lambda}$, and let (p^*, q^*, t^*) be the unique equilibrium.*

- (a) *If punishment for convicted agents is by imprisonment, then (p^*, q^*, t^*) is biased against lower types.*
- (b) *If punishment for convicted agents is by a simple fine, and agents are strictly risk averse (i.e. $u''(w) < 0 \forall w$), then (p^*, q^*, t^*) is biased against higher types.*

⁷Note that (9)–(9) describe the solution to a design problem which is not unlike the classical mechanism design problem. Similar to an incentive compatibility constraint, the unbiasedness requirement leads to a solution in terms of the slope of the design object, and similar to an individual rationality constraint, the fact that a certain deterrence effect must be created for the lowest type yields an initial condition.

In general, the shape of Δ depends on the shape of the both the utility function u and the penalty schedule ρ . In the following, we derive sufficient conditions for monotonicity of Δ , and hence for a bias in favor of higher or lower types, respectively. Let $R(w)$ denote the coefficient of relative risk aversion of u at w :

$$R(w) = -w \frac{u''(w)}{u'(w)}.$$

Further, let $\varepsilon(w)$ denote the income elasticity of ρ at w :

$$\varepsilon(w) = w \frac{\rho'(w)}{\rho(w)}.$$

We can then show that the following holds:

Theorem 6. *Suppose punishment for convicted agents is given by penalty schedule ρ . Suppose $0 < \lambda \leq \bar{\lambda}$, and let (p^*, q^*, t^*) be the unique equilibrium. Then (p^*, q^*, t^*) is biased against higher types if $R(w) > 1$ and $\varepsilon(w) \leq 1 \forall w$, and it is biased against lower types if $R(w) < 1$ and $\varepsilon(w) \geq 1 \forall w$.*

Hence, it is possible to extend the result of Theorem 4 (b): In the special case of a proportional fine, the income elasticity of the fine is zero, i.e. $\varepsilon(w) = 0$. Whether the equilibrium is biased against lower or higher types depends then *only* on whether the coefficient of relative risk aversion $R(w)$ is always below or above one.

It should be noted that, even when the equilibrium is biased, the judges's prejudice is consistent with the true likelihood that agents of various types commit crimes. Furthermore, in equilibrium agents with differing types take into account that there is a different chance of conviction if they commit a crime than for agents of other types. For example in the case of imprisonment, the prejudice against an agent is decreasing with his type, so agents with higher types know that they are more likely to avoid punishment when committing a crime. However, in equilibrium higher types are still less likely to commit crime than lower types (i.e. $q^*(w) > q^*(w')$ if $w > w'$).

We have thus established conditions for the treatment an individual receives from the courts to be dependent on their wealth. The story thus far suggests, however, that people of the same wealth, but that differ in other characteristics, should be treated the same. In the next section, we allow for income to be determined endogenously, and find that bias in the courts based on wealth can in fact lead to bias based on other characteristics.

4 Non-economic Types and Statistical Discrimination

In this section, we dispose of the assumption that agents differ in terms of their incomes w ex-ante. We instead assume that all agents are ex-ante alike with respect to economic characteristics such as their productivity, human capital, etc., which can influence a

person’s income. There are now $L \geq 2$ subgroups of the population, and membership in one group is merely a label $l \in \{1, \dots, L\}$ without economic significance. This label could be race, gender, nationality, or ethnicity.⁸ An agent’s group label is publicly observable, and all differences in the realized value of an agent’s income are the direct result of the bias held against the agent’s group in the justice system. If such real inter-group differences arise endogenously ex-post, then such outcomes are “sunspot equilibria.”

To justify the assumption that an agent’s income is a function of prejudice, note that it may simply be harder for a person who is assumed to be more prone to criminal behavior to find a job, or to find a high-paying job, resulting in lower income. This can be the case for a myriad of reasons. Criminal activity may adversely affect an individual’s productivity, or criminal activity may directly be targeted at the employer (e.g. stealing from a job site). Further, if there are training costs for new employees and employers expect that members of certain groups are more likely to be convicted of a crime and be sent to jail, then hiring a member of the disadvantaged group has a larger expected cost because this individual’s duration in employment is on average shorter. Alternatively, the qualifications required to be employed at a given wage can be costly to obtain and training must be paid by the individual in form of tuition or foregone income while at school. An individual’s decision to acquire these qualifications will depend on their expected return, which will be lower for those individuals who are likely to be incarcerated in the future. We do not focus on any particular such story, and instead take a reduced form approach. Specifically, we assume that there is a weakly decreasing and continuous function $v : [0, 1] \rightarrow [\underline{w}, \bar{w}]$ ($\underline{w} < \bar{w} < \infty$) which describes an agent’s income as a function of the prejudice held against the group to which he belongs.

Since the only observable difference between agents is that they belong to different subgroups of the population, replace the prejudice $p(w)$ against persons of income w by a prejudice p_l against persons who belong to group l . Similarly, replace $q(w)$ by q_l , and $t(w)$ by t_l . An equilibrium in this case is now a collection $(w_l^*, p_l^*, q_l^*, t_l^*)_{l \in \{1, \dots, L\}}$ such that for all $l \in \{1, \dots, L\}$,

$$p_l^* = 1 - Q(q_l^*), \quad (10)$$

$$q_l^* = [1 - F(t_l^*) - \lambda(1 - G(t_l^*))] \Delta_l, \quad (11)$$

$$\theta(p_l, t_l^*) = \alpha, \quad (12)$$

$$w_l^* = v(p_l^*), \quad (13)$$

where $\Delta_l = u(w_l^*) - u(w_l^* - \rho(w_l^*))$ is defined as before. This is essentially the same

⁸Of course, what we call “non-economic characteristics” can also be a-priori correlated with economically meaningful variables. A woman’s productivity as a lumberjack is presumably on average lower than a man’s, and a French person may on average be a better food critic than a British person. We assume such cases away in our model.

set of defining equations as (2)–(4), except that a fourth condition has been added (condition (13)). This condition states that the equilibrium income level of group l , w_l^* , must be consistent with the prejudice level p_l^* held against group l . We now say an equilibrium is *biased* if there exist $l, l' \in \{1, \dots, L\}$ such that $p_l^* > p_{l'}^*$. An equilibrium is *unbiased* if $p_1^* = \dots = p_L^*$.

Equilibria can be constructed from the points of intersection of two curves in p - w space. The first curve is the locus of all (p, w) pairs such that $p = p^*(w)$, where $p^*(w)$ is the equilibrium prejudice against an individual of income w , given by (2) of the simple model. The second one is the graph of the income function $v(p)$. Let

$$\mathcal{S} = \{(p, w) \in [0, 1] \times [\underline{w}, \bar{w}] : p = p^*(v(p))\}$$

be the set of all intersecting points of the two curves. An equilibrium in the model with endogenous income can then be constructed by assigning to each group $l \in \{1, \dots, L\}$ a point in \mathcal{S} , corresponding to the prejudice p_l^* against group l , and the income level w_l^* of its members. The conviction threshold applied to defendants from group l , t_l^* can then be computed from (12), and the decision threshold which agents from group l use, q_l^* , can be computed from (11).

If $\mathcal{S} \neq \emptyset$, then an equilibrium exists. The following states a sufficient condition for existence:

Lemma 7. *Regardless of the punishment used, if $0 < \lambda \leq \bar{\lambda}$ there exists an equilibrium in the model with endogenous income.*

If an equilibrium exists, there is always an unbiased equilibrium, as all groups can be assigned the same (p, w) -pair. Whenever \mathcal{S} contains more than one element, there are also biased equilibria, as any assignment of groups to (p, w) -pairs contained in \mathcal{S} represents an equilibrium in the extended model. Thus, if $|\mathcal{S}| > 1$, there is an equilibrium in which $p_l^* \neq p_{l'}^*$ for $l \neq l'$, and also $w_l^* \neq w_{l'}^*$, namely if $(p_l^*, w_l^*) \in \mathcal{S}$ and $(p_{l'}^*, w_{l'}^*) \in \mathcal{S}$. The following result mirrors Lemma 3.

Lemma 8. *Fix $\underline{w} > 0$ and $\bar{w} > \underline{w}$. Let \mathcal{D} be the set of all continuous, decreasing functions $v : [0, 1] \rightarrow [\underline{w}, \bar{w}]$. Suppose $\lambda \leq \bar{\lambda}$.*

- (a) *If Δ strictly decreases in w , or is constant, then for all $v \in \mathcal{D}$ there is a unique equilibrium, and this equilibrium is unbiased.*
- (b) *If Δ increases in w , there exists a non-empty set $\mathcal{D}_0 \subset \mathcal{D}$ such that for all $v \in \mathcal{D}_0$, a biased equilibrium exists.*

The next result follows then immediately from Lemma 8 (a proof is therefore omitted):

Theorem 9. *Fix $\underline{w} > 0$ and $\bar{w} > \underline{w}$. Let \mathcal{D} be the set of all continuous, decreasing functions $v : [0, 1] \rightarrow [\underline{w}, \bar{w}]$. Suppose $\lambda \leq \bar{\lambda}$.*

- (a) *If convicted offenders are punished by a simple fine, then for all $v \in \mathcal{D}$ there is a unique equilibrium, and this equilibrium is unbiased.*
- (b) *If convicted offenders are punished by imprisonment, there exists a non-empty set $\mathcal{D}_0 \subset \mathcal{D}$ such that for all $v \in \mathcal{D}_0$ a biased equilibrium exists.*

Figure 2 depicts the case of Theorem 9 (b), where $v \in \mathcal{D}_0$. It is worth mentioning that there is nothing “special” about the function v depicted. That is, this result does not rely on there being any points of tangency or discontinuities, and the set \mathcal{D}_0 could be loosely be regarded as “generic.”⁹ In general, there will be a relatively large number of functions $v \in \mathcal{D}_0$. With prison as punishment for convicted agents, the p^* -curve is increasing, and it is not hard to plot a continuous and decreasing v -curve which intersects the p^* -curve several times. In our example there are three intersections ($|\mathcal{S}| = 3$), so that one can construct equilibria with up to three different endogenous income levels. We depict an example with two groups: Group 1 earns a relatively high income w_1^* and faces a relatively low prejudice p_1^* . The opposite holds for group 2.

In contrast, Figure 3 shows the case of the same v -curve as before, but a simple fine is used instead of prison to punish convicted agents. This corresponds to Theorem 9 (a). In this case the p^* -curve is strictly increasing, and it is easy to see that there

⁹A formal definition of measure in function space is beyond the scope of this paper; however, see Nelson (1959) for details.

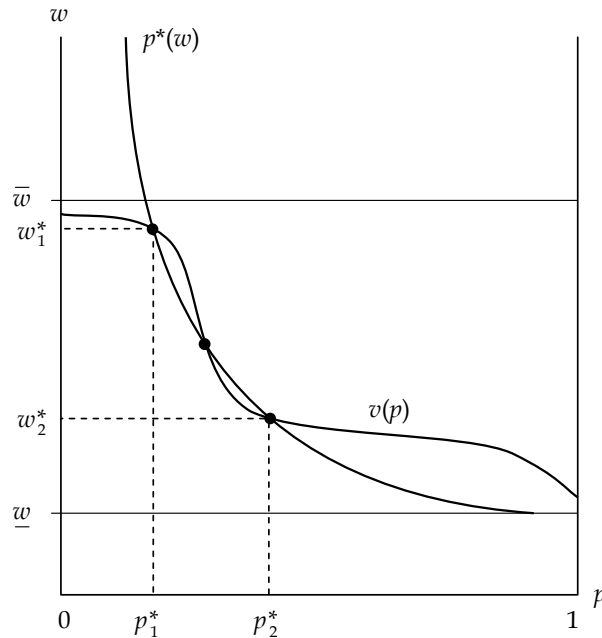


Figure 2: Biased equilibrium when punishment is by imprisonment

cannot be multiple intersections of p^* and v now. Thus the only equilibrium is an unbiased one, where both groups earn the same income and face the same prejudice.

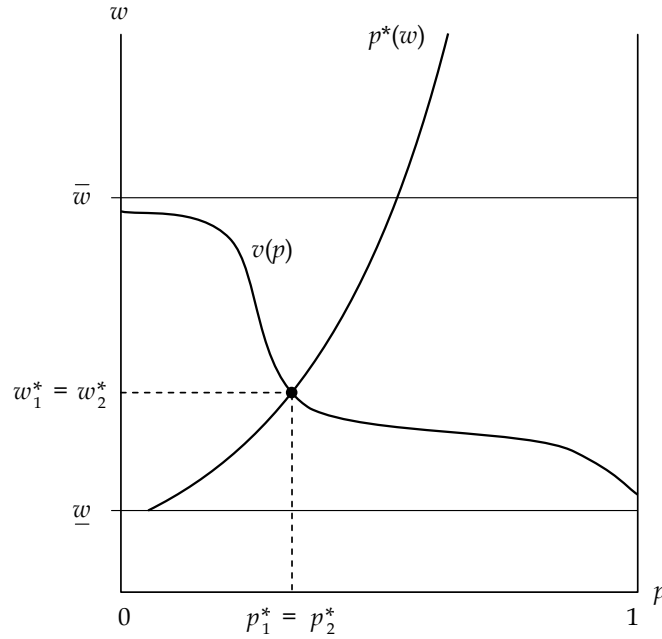


Figure 3: Unbiased equilibrium when simple fines are used

5 Extensions

In this section, we show how our model can be extended by introducing dynamics (Section 5.1) and the possibility of multiple offenses (Section 5.2). The merits of these extensions are described in detail below.

5.1 A Dynamic Link from Prejudice to Income

When punishment is by imprisonment, we have shown that there exists a possibility for discriminatory equilibria, where racial groups differ in terms of their crime rates, but also in terms of their incomes. This required there to be some feedback from the (perceived) proclivity to crime of an individual to the individual’s income. We modeled this feedback in “reduced form” by assuming there be a downward sloping function v that mapped an individual’s perceived crime rate to income. In equilibrium of such a model, both membership in a certain racial group, or an individual’s income, can then explain the likelihood that the individual is convicted of a crime. There is, however, empirical evidence (cited in the introduction) that race is a significant predictor of criminal activity *after* controlling for wages and a number of other economic variables.

In this section, we show that when introducing a temporal dimension to our model, this empirical observation can be reconciled. In particular, by using expected *lifetime* earnings as the relevant income variable, we show that the sanctions imposed by the courts on convicted criminals can be directly responsible for differences in lifetime earnings.

Suppose employers offer (voluntarily or by law) the same wage to everybody, as long as they possess the same economically relevant qualifications, such as academic degrees or technical diplomas. However, members of one group are more likely to spend some time of their lives in prison than others. They will hence receive the same wage as everybody else *when they work*, but over a stochastically shorter period of time. For simplicity, assume that only life sentences are given. Then individuals who face a strong prejudice on average will go to jail sooner—thus their expected lifetime earnings they stand to lose from the sentence are lower, compared to individuals who face a lesser prejudice. Consequently, they are less likely to be deterred by the threat of losing this expected future income, and more likely to commit crimes, making this an equilibrium again. The difference is now that a worker’s per-period wage income has no predictive power regarding criminal activity, but membership in racial groups has. Furthermore, it is not discrimination in the labor market, but the criminal justice system itself, that provides the causal link between prejudice and income. If wage differences exist that are due to differential treatment by employers, affirmative action that mandates non-discriminatory treatment *in the labor market* could help eliminate this problem (similar to the argument made by Coate and Loury (1993)). However, in the case we consider here, such policies will have no effect at best, as employers pay the same wage to all workers already.

We assume that agents live for infinitely many periods. At the beginning of each period, an agent will work and earn a *fixed* wage of y , unless he is in prison in which case he earns zero. There is no saving technology, and all income is consumed in the period it is earned. At the end of each period, the events described in Section 2 unfold: Agents observe their η -shocks and decide whether to commit the crime or not (the η -draws are assumed i.i.d. across agents and time), and must possibly stand trial. At the end of the period, the agent is then either a convicted criminal or not. If an agent is convicted (rightfully or wrongfully), he is removed from employment and sentenced to life in prison. This sentence represents a permanent reduction in income.¹⁰ Otherwise, the agent starts the next period as a working individual. In computing their expected lifetime earnings, agents apply a common discount factor $\beta < 1$ and do not include the future realizations of η . This is a behavioral assumption, of course, but unless η is literally regarded as the material benefit from a crime (for instance money stolen), this assumption seems not unreasonable. For example, one interpretation of η is that

¹⁰We expect that the results we derive continue to hold for sufficiently long but finite prison sentences, as the crucial feature of this penalty is not so much the duration of time over which it is applied, but that it reduces the agent’s lifetime income *to* a fixed level.

it represents the short-lived “kick” an individual gets from the crime. For offenses such as the consumption of illegal substances, it seems very natural to impose such a strong bias for the presence regarding the benefit η .

To solve this model, note that at the time the agent has observed the current period’s value of η and must decide whether or not to commit a crime, the expected lifetime consumption for an agent from the *next* period on, conditional on entering the next period as a free individual, is

$$v(p) = \frac{\beta}{1 - \beta\xi(q, t)}y, \quad (14)$$

where

$$\xi(q, t) = [1 - Q(q)]F(t) + Q(q)[1 - \lambda(1 - G(t))]$$

is the period-to-period “survival probability” associated with the tuple (q, t) . Since a given value for p pins down (q, t) via (2)–(4), we can write v as a function of p , as in (14). When an individual decides whether or not to commit a crime, the relevant utilities he must consider correspond to the prospect of earning zero from the next period onward (if convicted), or earning an expected continuation utility $v(p)$ from the next period onward. Hence we define

$$\Delta = v(p) - 0 = \frac{\beta}{1 - \beta\xi(q, t)}y.$$

We now show that this dynamic model can give rise to a discriminatory equilibrium, even though in this equilibrium every worker earns the same fixed wage y . To show that this can happen, we use an example using the following parameter values:

$$y = 1, \alpha = 0.95, \beta = 0.95, \lambda = 0.01, \eta \sim U[0, 5], F(t) = t^2, G(t) = 2t - t^2.$$

Note that this is not entirely in line with some of our previous assumptions; for instance $f(0) = 0$ and η is bounded in this example. However, those assumptions were made earlier because they were sufficient to ensure equilibria existed and had the properties we identified. They are not necessary, however, and our example illustrates that the same biased outcomes can also arise in other cases were the assumptions are violated.

We compute the $v(p)$ and $p^*(w)$ numerically. The result is plotted in Figure 4. One can see that there are in fact three intersections of the two curves. Thus, if there are two or more racial (or otherwise distinguishable) groups in the population, whose members all earn $y = 1$ when not incarcerated, biased equilibria can arise simply because two different groups can be “assigned” different prejudice-lifetime income pairs which correspond to the intersections in Figure 4.

This dynamic model provides an explanation for why some individuals choose to live a “life of crime,” and why this choice may be correlated with characteristics such as race. Consider an individual in the high crime/low lifetime income group. Each time he decides whether to commit a crime or not, he compares the benefit (η), which

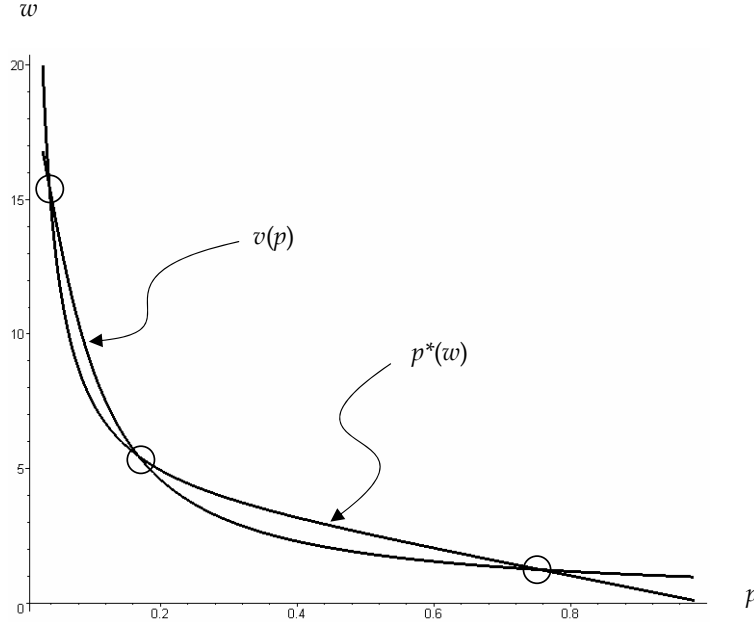


Figure 4: Numerical example ($v(p)$ is expected discounted lifetime income)

is distributed equally across the entire population, against the expected cost. The expected cost is that the individual (with some probability) loses his criminal career and goes to jail. However, continuing a life of crime is not a very enticing prospect either, because a career criminal expects to be jailed sooner or later anyways. Hence, such a person is less likely to be deterred by this prospect. The opposite holds for the choice to live a low-crime life, and the usual stereotyping argument can be made to sort individuals into different such equilibria, based on their race or other observable characteristics. The crucial aspect here is that, because race is observable and cannot be altered, it is not possible for individuals to break out of the high-crime equilibrium—they would be treated in an adverse manner by the courts even if they decided not to commit crime ever. Note that this result relies on the difference in the deterrence effect that long prison sentence have for members of the different subgroups. The same story could *not* be told in an alternative framework where instead of the cost of crime the benefit was different across groups (as would be the case, for example, if we focused on property crime and assumed that poorer individuals could gain more from stealing).

5.2 Multiple Offenses and Escalating Sanctions

In this section, we drop the interaction with the labor market. That is, we now consider individuals without regard to their income or wealth. As in Section 4, we consider $L \geq 2$ subgroups of the population, where members of these subgroups do not differ in any meaningful way. The timing is as before, with the exception that, when agents decide

whether to commit a crime, they do so knowing that there is a probability they may be (falsely) convicted of some *other* crime as well¹¹. Thus an individual may be convicted of zero, one or two crimes.

As before, we consider the utility costs of the penalties, although we no longer consider them as a function of the individual's wealth. Define the utility cost from the first penalty as Δ_1 , and the utility cost of the second penalty as Δ_2 . Thus, if an individual is convicted of a single crime, the penalty in terms of utility is Δ_1 , and if an individual is convicted of two crimes, the total utility cost is $\Delta_1 + \Delta_2$.

Let m_1 denote the probability that an individual is convicted of a crime that they did commit, and let m_0 denote the probability that the same individual is (wrongfully) convicted of a crime they did not commit. As before, if the agent does not commit a crime, there is a chance, m_0 , that the agent will still be convicted and pay a penalty. If the agent does commit a crime, there is now a chance, m_0m_1 , that they will be convicted of two crimes. The crime is committed if and only if the benefit from doing so exceeds its cost:

$$\eta > q \equiv m_1 [(1 - m_0)\Delta_1 + m_0\Delta_2].$$

As before, the probability of conviction following a crime is given by

$$m_1 = P[\theta(t) \geq \alpha | d = 1] = P[t \geq t | d = 1] = 1 - F(t),$$

and the probability of wrongful conviction is

$$m_0 = \lambda P[\theta(t) \geq \alpha | d = 0] = \lambda P[t \geq t | d = 0] = \lambda(1 - G(t)).$$

Given the difference in the expected cost of committing a crime, we now have that the tuple (p^*, q^*, t^*) constitutes an *equilibrium* if it solves the following system of equations:

$$p^* = 1 - Q(q^*) \tag{15}$$

$$q^* = [1 - F(t^*)] [(1 - \lambda(1 - G(t^*)))\Delta_1 + \lambda(1 - G(t^*))\Delta_2] \tag{16}$$

$$\theta(t^*) = \alpha \tag{17}$$

where equations (15) – (17) have the same interpretation as before. Note that only (16) is different, but it still is continuous with compact domain, and so the proof to the previous existence lemma applies here as well.

Lemma 10. *A rational expectations equilibrium exists.*

Before, we found that the equilibrium was unique as long as q^* was decreasing in t^* . Examining equation (16) yields

$$\begin{aligned} \frac{\partial q}{\partial t} &= -f(t) [(1 - \lambda(1 - G(t)))\Delta_1 + \lambda(1 - G(t))\Delta_2] + [1 - F(t)] [\lambda g(t) (\Delta_1 - \Delta_2)] \\ &= -f(t)\Delta_1 + \lambda [[1 - G(t)] f(t) + [1 - F(t)] g(t)] [\Delta_1 - \Delta_2] \end{aligned} \tag{18}$$

¹¹This other crime may be either a crime that was actually committed by someone else, or another activity undertaken by the individual that has been incorrectly interpreted as a criminal act.

and we have proved the following lemma:

Lemma 11. *The rational expectations equilibrium is unique when $\Delta_2 > \Delta_1$.*

When the equilibrium is not unique, stereotyping equilibria are possible, as was depicted previously in Figure 2. As long as second penalties are more severe in their utility costs, crime rates are uniquely determined across subgroups and there will be no stereotyping equilibrium. It should be noted that, with respect to fines, the utility cost of punishment is increasing ($\Delta_2 > \Delta_1$) even when the amount of the fine is independent of the number of convictions because of diminishing marginal utility. However, with prison, the second conviction is likely to be less severe. Possible reasons for this include discounting (a second penalty occurs after the first has been served) and a general loss of social status, and perhaps employment and marital status as well, that occurs after the first conviction. As a result, if incarceration is used as punishment, prison terms that increase in the number of convictions would be desirable. Note that this provides an additional rationale for increasing (or at least non-decreasing) penalties, to complement the work of Emons (2006) and Stigler (1970). Note that even when the equilibrium is unique, it is still possible for groups that vary in non-economic characteristics (such as race) to differ in their crime rates through the mechanism described in section 4.

6 Conclusion

We developed a model that tied jury prejudice, the decision to commit crime, and conviction standards into a single framework. Within this framework, we were able to characterize the effects of different penalty forms, such as prison sentences or monetary sanctions, on beliefs and thus also on behavior. Making prejudice (i.e. beliefs) endogenous in the model allowed us to develop from it a theory of discrimination based on non-economic characteristics such as race. We now conclude the paper with a couple of remarks, concerning the implications of the paper and possible areas of further research.

First, our results have implications for the design, or reform, of criminal justice systems. We have compared the different effects of prison and monetary sentences on prejudice; the former being more likely to invite stereotyping than the latter. We do not intend to give a thorough survey of the advantages and disadvantages of fines vs. prison sentences here. Nevertheless, we think our model has something to say about the various forms of penalizing felonies such as minor drug offenses. In the U.S. such crimes are routinely punished by incarceration, while European countries tend to use fines. The example of drug offenses is particularly interesting, as drug-related crime is responsible for a large fraction of the U.S. prison population¹². In addition, those

¹²About 55% of Federal inmates in 2003 and 25% of state level inmates were incarcerated for drug-related offenses. Source: Bureau of Justice Statistics Prisoners in 2004, NCJ 210677, October 2005.

convicted for drug-related offenses are disproportionately Black or Hispanic¹³. If the possibility of statistical discrimination is a concern to those who draft sentencing laws, then the relationship between different forms of penalties and the effects they have on crime and prejudice should be taken seriously.

Second, note that our model (with prison sentences) produced two related phenomena: In the static version, sunspot equilibria could emerge in which people were treated differently by the courts based on their race. In such equilibria, the group which faced a less favorable judicial prejudice was also ex-post economically disadvantaged. In the dynamic version of Section 5.1, we constructed an example that possessed a similar sunspot equilibrium; however, the group with the less favorable prejudice was disadvantaged only in terms of their lifetime earnings, and not in terms of their per-period earnings when at work. Obviously, a more elaborate model can be conceived in which both aspects arise: The disadvantaged racial group has on average lower wages than the advantaged one, for the reasons mentioned in the text, but race has residual predictive power of criminal activity. This pattern exists very clearly in the data (see the references cited in the introduction), and we regard the lifetime income model as a promising theoretical framework in which to explore this issue further.

Finally, we have shown that the model can be extended to multiple offenses and more complex penalty schedules which are functions of the number of previous convictions. Examining the effects of escalating sanctions in the dynamic model is likely to produce interesting results. One rationale for escalating sanctions is that convicted offenders can be of two types: Those who are corrigible (and deserve a “second chance”) and those who are not (and from which the public should be protected). A person convicted of multiple offenses is more likely to belong to the second category; hence the increasing sentences for repeat offenders.¹⁴ It appears that in a dynamic model, where crime decisions are based on lifetime expected income, the number of prior convictions could serve a similar role as race in our model: If persons with prior convictions face an unfavorable prejudice in the legal system, then they may be more likely to commit crimes, which may make them appear to be incorrigible. Exploring this possibility is another question left for future research.

¹³Blacks and Hispanics represented 24% and 23% of those convicted for drug-related offenses in 2003, compared to 14% Whites. Source: Prisoners in 2005, NCJ 215092, November 2006.

¹⁴An alternative justification for escalating sanctions is given in Emons (2006).

Appendix

Proof of Lemma 1. To prove an equilibrium exists, we make a standard fixed point argument. Fix any w and define three maps,

$$\begin{aligned}\mathcal{T}_1 & : q \rightarrow p : [0, \infty) \rightarrow [0, 1], \\ \mathcal{T}_2^w & : t \rightarrow q : [0, 1] \rightarrow [0, \infty), \\ \mathcal{T}_3 & : p \rightarrow t : (0, 1) \rightarrow [0, 1]\end{aligned}$$

by (2), (3) (given w), and (4), respectively; these are all given in Section 2.5. Note that \mathcal{T}_1 , \mathcal{T}_2^w and \mathcal{T}_3 are continuous on their respective domains. \mathcal{T}_3 is well-defined through (4) on $(0, 1)$ only; however it can be extended continuously to $[0, 1]$ by setting $\mathcal{T}_3(0) = 1$ and $\mathcal{T}_3(1) = 0$. Further, as \mathcal{T}_2^w is continuous on a compact domain, its image is bounded. We can hence restrict the range of \mathcal{T}_2^w , as well as the domain of \mathcal{T}_1 , to $[0, \hat{q}(w)]$ for sufficiently large $\hat{q}(w)$. Now define a new map

$$\mathcal{T}^w : [0, 1] \times [0, \hat{q}(w)] \times [0, 1] \rightarrow [0, 1] \times [0, \hat{q}(w)] \times [0, 1]$$

by

$$\mathcal{T}^w(p, q, t) = (\mathcal{T}_1(q), \mathcal{T}_2^w(t), \mathcal{T}_3(p)).$$

Since \mathcal{T}^w maps a compact subset of \mathbb{R}^3 into itself, we can apply Brouwer's fixed point theorem to show, for given w , there exists $(p^*(w), q^*(w), t^*(w))$ such that

$$(p^*(w), q^*(w), t^*(w)) = \mathcal{T}^w(p^*(w), q^*(w), t^*(w));$$

thus it solves (2)–(4) simultaneously. Since such a fixed point can be constructed for each w independently, an equilibrium as defined above exists. \square

Proof of Lemma 2. To prove uniqueness, let $\lambda \leq \bar{\lambda}$ and suppose there are two equilibria, $(p^*, q^*, t^*) \neq (\tilde{p}^*, \tilde{q}^*, \tilde{t}^*)$. Thus there exists w such that $q^*(w) \neq \tilde{q}^*(w)$ (otherwise $p^*(w) = \tilde{p}^*(w) \forall w$ by (2), which implies $t^*(w) = \tilde{t}^*(w) \forall w$, by (4), but then the equilibrium would be unique). So suppose, without loss of generality, that $q^*(w) > \tilde{q}^*(w)$ for some w . Condition (2) then implies $p^*(w) < \tilde{p}^*(w)$, and using (4) we have $t^*(w) > \tilde{t}^*(w)$. If $\lambda \leq \bar{\lambda}$ then using (5) we get $q^*(w) \leq \tilde{q}^*(w)$, a contradiction. Hence the equilibrium is unique if $\lambda \leq \bar{\lambda}$. \square

Proof of Lemma 3. Let $w > w'$ and suppose $\Delta(w) > \Delta(w')$. If $p(w) \geq p(w')$, then by condition (2), $q(w) \leq q(w')$. However, by condition (4) and the fact that f/g increases, we have $t(w) \leq t(w')$. Thus if $\Delta(w) > \Delta(w')$ then (5)–(6) imply $q(w) > q(w')$, which is a contradiction, and therefore $p(w) < p(w')$. From (4) it follows then that $t(w) > t(w')$, and from (2) it follows that $q(w) > q(w')$. Exactly the opposite argument can be made when $w > w'$ and $\Delta(w) < \Delta(w')$. Finally, when Δ is a constant then (3) is independent of w so that p^* , q^* , and t^* are constant and hence constitute an unbiased equilibrium. \square

Proof of Theorem 4. If $u''(w) = 0$, then $\Delta'(w) = u'(w) - u'(w - \delta) = 0$, and Lemma 3 (iii) implies (a). If $\rho(w) = \gamma w$ and $u(w) = a \ln w + b$, then $\Delta'(w) = a/w - a/w = 0$, and Lemma 3 (iii) implies (b) as well. \square

Proof of Theorem 5. In case of imprisonment, $\Delta(w) = u(w) - u_0$, which is strictly increasing in w since $u'(w) > 0$. Applying Lemma 3 (ii) yields (a). In case of a fine, $\Delta(w) = u(w) - u(w - \delta)$, which is strictly decreasing in w if $u''(w) < 0 \forall w$. Applying Lemma 3 (i) yields (b). \square

Proof of Theorem 6. Suppose $R(w) > 1$ for all w , or equivalently

$$-\gamma w \frac{u''(\gamma w)}{u'(\gamma w)} > 1 \quad \forall \gamma,$$

and thus

$$\frac{\partial}{\partial \gamma} [\gamma u'(\gamma w)] = u'(\gamma w) + \gamma w u''(\gamma w) < 0 \quad \forall \gamma. \quad (19)$$

Let $\mu(w) = w - \rho(w)$ be the income left to the individual after the fine $\rho(w)$. Since $\mu(w) < w$, (19) implies

$$\frac{\mu(w)}{w} u'(\mu(w)) = u'(w) - \int_{\mu(w)/w}^1 \gamma u'(\gamma w) d\gamma > u'(w). \quad (20)$$

Suppose now that $\varepsilon(w) \leq 1 \forall w$: $w\rho'(w)/\rho(w) \leq 1$. Multiplying this inequality by $\rho(w)/w$ yields $\rho'(w) \leq \rho(w)/w$, and expressing the fine as $\rho(w) = w - \mu(w)$ we get $\mu'(w) \geq \mu(w)/w$. Then by (20)

$$\begin{aligned} \Delta'(w) &= \frac{\partial}{\partial w} [u(w) - u(\mu(w))] \\ &= u'(w) - \mu'(w)u'(\mu(w)) \\ &\leq u'(w) - \frac{\rho(w)}{w} u'(\mu(w)) < 0. \end{aligned}$$

By Lemma 3 (i), therefore, the equilibrium is biased against higher types. Analogous steps can be repeated for $R(w) < 1$ and $\varepsilon(w) \geq 1$, in which case $\Delta'(w) > 0$ and the equilibrium is biased against lower types by Lemma 3 (ii). \square

Proof of Lemma 7. If $\mathcal{S} \neq \emptyset$, then an unbiased equilibrium exists, as argued in the text. We have to show that $\mathcal{S} \neq \emptyset$. Note that v is a continuous function mapping $p \in [0, 1]$ to $w \in [\underline{w}, \bar{w}]$. From the simple model with exogenous income differences we borrow the map p^* , a correspondence which assigns to income levels $w \in [\underline{w}, \infty)$ prejudice levels $p \in [0, 1]$. If $\lambda \leq \bar{\lambda}$, $p^*(w)$ is single-valued by Lemma 1. As Δ is continuous in w regardless of the punishment used, w enters the mapping \mathcal{T} defined in the proof of Lemma 1 continuously, which implies that p^* is upper-hemicontinuous. But then p^* can equivalently be expressed as a continuous function from $[\underline{w}, \infty)$ to $[0, 1]$. This implies that in $[0, 1] \times [\underline{w}, \bar{w}]$, the graphs of these functions intersect at least once, so $\mathcal{S} \neq \emptyset$. \square

Proof of Lemma 8. In the proof of Theorem 7 we already established that p^* is a continuous function from $[\underline{w}, \infty) \rightarrow [0, 1]$. We first prove (a). Lemma 3 (i), p^* strictly increases for strictly decreasing Δ . Therefore, for each $v \in \mathcal{D}$ there is exactly one $w \in [0, 1]$ such that $v(p^*(w)) = w$. If Δ is constant, then by Lemma 3 (iii) p^* is constant. It will hence become a vertical line in p - w space, which is intersected by any decreasing, continuous v exactly once. Hence $|\mathcal{S}| = 1$ and a unique, unbiased equilibrium exists. To prove (b), note that by Lemma 3 (ii) p^* strictly decreases for strictly increasing Δ . Therefore, there exists a non-empty set of continuous, decreasing functions $v : [0, 1] \rightarrow [\underline{w}, \bar{w}]$ for which the following holds: There exists $w_1, w_2 \in [\underline{w}, \bar{w}]$, $w_1 \neq w_2$, such that $v(p^*(w_1)) = w_1$ and $v(p^*(w_2)) = w_2$. For all such v , $|\mathcal{S}| > 1$ and a biased equilibrium exists. \square

References

- ALEXEEV, M. AND J. LEITZEL (2004): “Racial Profiling,” mimeo, Indiana University.
- ARROW, K. (1973): “The Theory of Discrimination,” in: O. Ashenfelter and A. Rees (eds.): *Discrimination in Labor Markets*, Princeton University Press, 3–33.
- AUMANN, R. (1987): “Correlated Equilibrium as an Expression of Bayesian Rationality,” *Econometrica*, **55**, 1–18.
- BECKER, G. (1957): *The Economics of Discrimination*, University of Chicago Press.
- BECKER, G. (1968): “Crime and Punishment: An Economic Approach,” *Journal of Political Economy*, **76**, 169–217.
- BJERK, D. (2006): “The Effects of Segregation on Crime Rates,” mimeo, McMaster University.
- BJERK, D. (2007): “Racial Profiling, Statistical Discrimination, and the Effect of a Colorblind Policy on the Crime Rate,” *Journal of Public Economic Theory*, forthcoming.
- BURKE, A. (2007): “Neutralizing Cognitive Bias: An Invitation to Prosecutors,” mimeo, Hofstra University School of Law.
- CARTWRIGHT, E. AND M. WOODERS (2006): “Conformity, Correlation, and Equity,” mimeo, Vanderbilt University.
- CASS, D. AND K. SHELL (1983): “Do Sunspots Matter?” *Journal of Political Economy*, **91**, 193–227.

- COATE, S. AND G. LOURY (1993): “Will Affirmative-Action Policies Eliminate Negative Stereotypes,” *American Economic Review*, **83**, 1220–1240.
- EMONS, W. (2006): “Escalating Penalties for Repeat Offenders,” *International Review of Law and Economics*, forthcoming.
- FORGES, F. (1986): “An Approach to Communication Equilibria,” *Econometrica*, **54**, 1375–1385.
- GEORGAKOPOULOS, N. (2004): “Self-fulfilling Impressions of Criminality: Unintentional Race Profiling,” *International Review of Law and Economics*, **24**, 169–190.
- KNOWLES, J., N. PERSICO, AND P. TODD (2001): “Racial Bias in Motor-Vehicle Searches: Theory and Evidence,” *Journal of Political Economy*, **109**, 203–229.
- KRIVO, L. AND R. PETERSON (1996): “Extremely Disadvantaged Neighborhoods and Urban Crime,” *Social Forces*, **75**, 619–650.
- LEVITT, S. (1997): “Incentive Compatibility Constraints as an Explanation for the Use of Prison Sentences Instead of Fines,” *International Review of Law and Economics*, **17**, 179–192.
- MORRIS, N. AND M. TONRY (1990): *Between Prison and Probation*, Oxford University Press, Oxford.
- NELSON, E. (1959): “Regular Probability Measure on Function Space,” *Annals of Mathematics*, **69**, 630–643.
- PERSICO, N. (2002): “Racial Profiling, Fairness, and Effectiveness of Policing,” *American Economic Review*, **92**, 1472–1497.
- PHELPS, E. (1972): “The Statistical Theory of Racism and Sexism,” *American Economic Review*, **62**, 659–661.
- POLINSKY A. AND S. SHAVELL (1984): “The Optimal Use of Fines and Imprisonment,” *Journal of Public Economics*, **24**, 89–99.
- POSNER, R. (1992): *Economic Analysis of Law*, 4th Edition, Little, Brown and Company, Boston.
- RAPHAEL, S. AND R. WINTER-EBMER (2001): “Identifying the Effect of Unemployment on Crime,” *Journal of Law and Economics*, **44**, 259–283.
- SAH, R. (1991): “Social Osmosis and Patterns of Crime,” *Journal of Political Economy*, **99**, 1272–1295.

STIGLER, G. (1970): “The Optimum Enforcement of Laws,” *Journal of Political Economy*, **78**, 526–536.

TRUMBULL, W. (1989): “Estimations of the Economic Model of Crime Using Aggregate and Individual Level Data,” *Southern Economic Journal*, **56**, 423–439.

VERDIER, T. AND Y. ZENOU (2004): “Racial Beliefs, Location, and the Cause of Crime,” *International Economic Review*, **45**, 731–760.