What Marginal Outcome Tests Can Tell Us About Racially Biased Decision-Making*

Peter Hull†

July 2021

Abstract

Marginal outcome tests compare the expected effects of a decision on individuals who are of different races but at the same indifference point of the decision-maker. I present a simple formalization of how such tests can detect racial bias, defined as a deviation from accurate statistical discrimination. Namely, the tests can reject that the decision-maker ranks individuals according to some accurate prediction of a mandated outcome, given some unspecified race-inclusive information set. The frontier of marginal effects can furthermore rule out canonical taste-based discrimination. I relate this analysis to other interpretations of marginal outcome tests, other notions of racial discrimination, and recent identification strategies.

*I thank David Arnold, Ivan Canay, Will Dobbie, Nicolás Grau, Jim Heckman, Conrad Miller, Magne Mogstad, Jack Mountjoy, Jon Roth, Damián Vergara, and Crystal Yang for many helpful comments and conversations. Jerray Chang provided excellent research assistance.

†Brown University and NBER. Email: peter_hull@brown.edu

1 Introduction

Marginal outcome tests compare the average effects of a binary decision D_i on an outcome Y_i between individuals i who are of different races (or another protected characteristic) but at the same indifference point of the decision-maker. The decision-maker has a narrow stated or legal mandate pertaining to the outcome. Such tests are increasingly conducted in the criminal justice setting (e.g. Arnold et al., 2018, 2020a; Marx, 2018; Feigenberg and Miller, 2020; Grau and Vergara, 2021), where the decision-maker (e.g. a judge or police officer) takes an action D_i (e.g. the granting of pretrial release or the searching of a vehicle) that affects a given measure of criminal activity Y_i (e.g. pretrial misconduct or the finding of contraband). Motivated by the classic theory of Becker (1957), the finding of racially disparate marginal outcomes in such settings is often taken as an indication of biased decision-making.

Despite their recent proliferation, however, it is not always clear what marginal outcome tests reveal about racial bias, and under what conditions. In contrast to an earlier literature which specifies and estimates complete decision-making models (e.g. Knowles et al., 2001; Antonovics and Knight, 2009; Persico and Todd, 2006; Anwar et al., 2012), recent empirical applications leverage quasi-experimental variation to identify and compare marginal outcomes while assuming minimal structure. An advantage of this quasi-experimental approach is a robustness to alternative models, requiring only the identification of race-specific marginal treatment effects (MTEs). But how such MTE comparisons can be interpreted in terms of racial bias, and under what assumptions, can be unclear (Canay et al., 2020).

This paper presents a simple theoretical framework for interpreting marginal outcome tests in terms of racially biased decision-making, without specifying a decision-making model. The only maintained assumption is one sufficient to define the marginal outcome test: that the decision-maker acts according to some subjective ranking of individuals by their perceived appropriateness for the treatment. The existence of such a decision rule defines a set of race-specific marginal treatment effect frontiers, which capture the relationship between the decision-maker's subjective ranking and the objective effects on the outcome. I do not assume this ranking arises from a particular optimization problem (as in Knowles et al. (2001)) nor that it satisfies certain separability conditions (as in Canay et al. (2020)).

I first show that the marginal outcome test can generally detect a notion of racial bias that is defined as a deviation from canonical statistical discrimination (Phelps, 1972; Arrow, 1973; Aigner and Cain, 1977). Formally, I show the test can reject that the decision-maker's ranking of individuals reflects some accurate prediction of outcomes, given by some subjective information set that includes the individual's race. Forms of racial bias that drive test rejections include canonical taste-based discrimination, as in Becker (1957), as well as biased beliefs or stereotypes, as in Bordalo et al. (2016) and Bohren et al. (2020). I show that while

marginal outcome tests generally cannot differentiate between these alternative hypotheses, the slopes of the MTE frontiers can reject canonical taste-based discrimination.[1] I discuss how rejecting accurate statistical discrimination and canonical taste-based discrimination can be relevant to policy, even though the decision-making model is unspecified.

I then relate this definition of racial bias, and interpretation of marginal outcome tests, to existing definitions and analyses. In contrast to the recent analysis of Canay et al. (2020), I find that marginal outcome tests can be informative about racial bias without an extended Roy model restriction (Heckman and Vytlacil, 2007; D'Haultfœuille and Maurel, 2013). This different conclusion stems from my definition of racial bias as a deviation from accurate statistical discrimination.[2] In contrast, Canay et al. (2020) define racial bias with respect to a particular decision-maker information set, which can be restrictive or difficult to specify in practice. I further discuss how this paper's definition of bias differs from other notions of discrimination, as estimated recently by Arnold et al. (2020a) and Feigenberg and Miller (2020), and how it aligns with the notion of algorithmic "calibration" in the computer science literature (e.g. Barocas and Selbst, 2016; Kleinberg et al., 2017a; Chouldechova, 2017).

I conclude with a discussion of how marginal outcome tests might be conducted. Identification of marginal effects can follow from an experimental intervention, in which decision-makers are induced to increase or decrease their treatment rates among a random set of white and Black individuals. Some quasi-experimental applications are based on this idealized within-agent comparison (e.g. Feigenberg and Miller (2020); see also Grau and Vergara (2021) for an observational approach in the same spirit). In other settings, identification may come from observing multiple as-good-as-randomly assigned decision-makers (e.g. Arnold et al., 2018, 2020a; Marx, 2018). Across-agent variation can be used to estimate within-agent margins under restrictions on the heterogeneity in MTE frontiers: either conventional monotonicity assumptions that impose MTE uniformity, or alternative restrictions that tractably parameterize MTE heterogeneity (Arnold et al., 2020a).

The remainder of this paper is organized as follows. Section 2 presents the setting and defines the marginal outcome test. Section 3 formalizes how this test can detect racial bias, and how canonical taste-based discrimination can be further detected. Section 4 links this analysis to others in the literature. Section 5 discusses identification. Section 6 concludes.

[1] As discussed below, this relates to results in Marx (2018), Canay et al. (2020), and Gelbach (2021).

[2] This definition appears to align with the motivation for some recent empirical applications of the marginal outcome test (e.g Arnold et al., 2018; see also Arnold et al., 2020b), though as Canay et al. (2020) note these motivations are not always precisely formalized. I emphasize that the purpose of this paper is not to promote any particular interpretation of the marginal outcome test as the intended one in a given application. Rather, my goal is to provide a novel formalization that is in line with the canonical theory of Becker (1957), Phelps (1972), Arrow (1973), and Aigner and Cain (1977); has potential policy consequences; and requires minimal assumptions on the decision-making model and information set.

2 Setting

I consider a population of individuals i differentiated by an observable characteristic $R_i \in \{w, b\}$. For concreteness I refer to R_i as individual i's race (either "white" or "Black") though it could of course index another characteristic such as age or sex. A decision-maker (for concreteness, a "judge") observes individual race and some other information S_i. Her binary decision, or "treatment," is then given by $D_i \in \{0, 1\}$. The judge has some mandate to base treatment decisions around their likely effect on an individual outcome Y_i. Letting Y_{i1} and Y_{i0} denote individual i's potential outcome with and without treatment, these treatment effects are written $Y_i^* = Y_{i1} - Y_{i0}$. Throughout I assume that Y_i is an adverse outcome which the judge seeks to minimize. Otherwise Y_i can be interpreted as the negation of her mandate.

A concrete example of the setting is given by Arnold et al. (2018), who study pretrial release decisions. Here $D_i = 1$ indicates that the pretrial judge releases a defendant before trial, with $D_i = 0$ when a defendant is detained. The legal mandate of such judges is typically narrow: to control the rate of pretrial misconduct (often defined as either a failure to subsequently appear in court or the participation in further crimes), an adverse outcome measured by a binary Y_i. In this setting Y_i^* is a binary latent variable. Since a detained defendant cannot conduct pretrial misconduct, $Y_{i0} = 0$ for all i. Thus the effect of release is simply a defendant's potential for misconduct if released: $Y_i^* = Y_{i1} \in \{0, 1\}$. Another setting with binary Y_i^* is given by Feigenberg and Miller (2020), who study traffic stop decisions D_i. Here $Y_i = 1$ indicates the failure to detect illegal contraband. Since contraband cannot be detected among non-stopped drivers, we again have $Y_{i0} = 0$ and $Y_i^* = Y_{i1} \in \{0, 1\}$ indicates latent contraband status. I use the binary Y_i^* case to strengthen and build intuition for some results, though the main analysis holds for Y_i^* with generic support.[3]

Throughout the paper I maintain a single assumption on the judge's decision-making process: that it can be represented by a subjective ranking of individuals by their appropriateness for treatment, with some individuals of each race marginal in the decision:

Assumption 1. There exists a random variable U_i such that $U_i \mid R_i$ is continuously distributed, judge treatment decisions can be represented as $D_i = \mathbf{1}[U_i \leq 0]$, and treatment rates $\pi_r \equiv Pr(U_i \leq 0 \mid R_i = r)$ are strictly between zero and one for each $r \in \{w, b\}$.

I refer to U_i as the judge's "ranking" of individuals. The existence of this latent variable is without much conceptual loss, saying only that one could in principle order individuals in the population according to the judge's perception of their qualification for treatment (with the threshold for qualification normalized to zero). Individuals with $U_i > 0$ are those the

[3] In the binary case it may be more natural to think of Y_i^* as a latent state instead of a treatment effect. This is more in line with the earlier model-based approach (e.g. Knowles et al., 2001) and may be more natural for some identification strategies (e.g. Grau and Vergara, 2021).

judge thinks are unqualified for treatment while those with $U_i \leq 0$ are deemed qualified. The continuity condition implies that the judge is in principle indifferent to treating some individuals of each race. Finally, Assumption 1 requires that some individuals of each race are both treated and untreated.

We first use Assumption 1 to define the marginal outcome test:

Definition 1. Let $\beta = E[Y_i^* \mid U_i = 0, R_i = w] - E[Y_i^* \mid U_i = 0, R_i = b]$. The null hypothesis of the *marginal outcome test* is $\beta = 0$.

The two terms of β give the expected treatment effects of the white or Black individuals that the judge is indifferent to treating. The test asks whether these marginal effects differ.

Applications of the marginal outcome test often draw on the theory of marginal treatment effects (MTEs; Heckman and Vytlacil, 2005). Our first result establishes the equivalence of β and the difference in two race-specific MTEs:

Lemma 1. There exists a random variable W_i with $W \mid R_i \sim \mathrm{U}(0,1)$ and $D_i = \mathbf{1}[W_i \leq \pi_{R_i}]$. $\beta = \mu(\pi_w, w) - \mu(\pi_b, b)$, where $\mu(m, r) = E[Y_i^* \mid W_i = m, R_i = r]$.

The proof to Lemma 1, as with other results in this paper, is given in the appendix. It defines W_i by applying a simple conditional probability integral transform to $U_i \mid R_i$. Under some assumptions the MTE frontiers $\mu(m, r)$, and thus β, may be identified from the observed decisions of judges.[4] I discuss such identification strategies in Section 5, but focus first on what $\beta \neq 0$ could reveal about racially biased decision-making if it were known.

3 Main Results

To analyze β, I first define a notion of racial bias that contrasts with accurate statistical discrimination. Such bias can arise either from the judge's racial preferences or from biased "beliefs" (e.g. stereotypes). I then show how the marginal outcome test can detect this notion of bias, and how other information on the race-specific MTE curves can furthermore rule out canonical taste-based discrimination as a driver.

3.1 Defining Racial Bias

To define racial bias, I first define accurate statistical discrimination. Statistical discrimination models were originally proposed by Phelps (1972), Arrow (1973), and Aigner and

[4]Importantly, here we allow U_i and thus $\mu(m, r)$ to be judge-specific. Differences in the rankings of individuals across judges will generally violate the conventional monotonicity assumption from using judge assignment to instrument for D_i and measure MTEs. We return to this issue in the identification discussion.

Cain (1977) to explain persistent differences in an employer's treatment of equally productive white and Black workers, in the absence of the preference-based racial bias considered by Becker (1957). The statistical discrimination models show how such differences can arise when employers act in a race-neutral way on accurate predictions of worker productivity, given by the worker's observed race and a noisy "signal" of their productivity.

Consistent with this theoretical literature, I define accurate statistical discrimination by the existence of a race-inclusive information set that can rationalize the judge's ranking of individuals for treatment as being based on an accurate prediction of her mandate Y_i^* (in place of labor market productivity). I define racial bias by the lack of such an information set: that is, by the inability to rationalize the judge's ranking in terms of an accurate prediction of individual treatment effects, based on race and some non-race signal S_i:

Definition 2. The judge can be said to be engaged in *accurate statistical discrimination* if there exists an S_i such that $U_i = E[Y_i^* \mid R_i, S_i] - t$ for some t. If not, we say her decisions exhibit *racial bias*.

Here t denotes a threshold the judge puts on the accurate predictions $E[Y_i^* \mid R_i, S_i]$ when engaged in accurate statistical discrimination.[5]

In the case of binary Y_i^*, like the pretrial release setting of Arnold et al. (2018), accurate statistical discrimination and racial bias have intuitive representations in terms of rational expected utility maximization:

Lemma 2. If $Y_i^* \in \{0, 1\}$ then a judge's decisions exhibit racial bias if and only if there is no S_i such that

$$D_i \in \arg\max_{D(\cdot)\in\mathbf{D}} E\left[\mathcal{U}(Y_i, D(R_i, S_i)) \mid R_i, S_i\right], \tag{1}$$

for some $\mathcal{U}(y, d) : \{0, 1\}^2 \to \mathbb{R}$ with $\mathcal{U}(1, 1) - (0, 1) < 0$, where $\mathbf{D}$ is the set of decision rules $D(r, s) \in \{0, 1\}$ for $r \in \{w, b\}$ and s in the support of S_i.

This result follows from the fact that when Y_i^* is binary it has only one moment (the mean) that is relevant to the judge's mandate. The existence of a utility function $\mathcal{U}(y, d)$ that rationalizes her behavior along with some signal S_i is thus equivalent to her decisions being based on an accurate mean prediction of $E[Y_i^* \mid R_i, S_i]$. The $\mathcal{U}(1, 1) - (0, 1) < 0$ condition restricts the set of possible utility functions to those generating negative payoffs to the adverse outcome. The appendix proof shows that a similar result follows for general Y_i^* if we restrict to the judge maximizing the expectation of a utility function that is linear (and

[5]Note that since a judge's decision-making process is only unique up to an increasing transformation of U_i, racial bias more generally implies there is no S_i, t, and increasing $f(\cdot)$ such that $U_i = f(E[Y_i^* \mid R_i, S_i] - t)$. I abstract away from such $f(\cdot)$ to keep the notation and discussion as simple as possible.

monotone) in y: that is, if we assume the judge is risk-neutral in evaluating the effects of her decisions on the outcome.

It is also natural to define a notion of taste-based discrimination, inspired by Becker (1957). In this scenario the judge accurately predicts Y_i^* according to some race-inclusive information set, but uses race-specific thresholds for treatment reflecting racial preferences:

Definition 3. The judge can be said to be engaged in *canonical taste-based discrimination* if there exists an S_i such that $U_i = E[Y_i^* \mid R_i, S_i] - t(R_i)$ for some $t(w) \neq t(b)$.

Thus, a judge engaged in canonical taste-based discrimination has a decision rule of

$$D_i = \mathbf{1}[E[Y_i^* \mid R_i, S_i] \leq t(R_i)], \tag{2}$$

with different thresholds $t(r)$ for white and Black individuals. We may think of $t(w) > t(b)$ as taste-based discrimination against Black individuals and $t(b) > t(w)$ as taste-based discrimination against white individuals. I show below that a judge who can be said to be engaged in canonical taste-based discrimination cannot be said to be engaged in accurate statistical discrimination, aligning Definitions 2 and 3 with the original theoretical division of Becker (1957) and Aigner and Cain (1977).

Like statistical discrimination, canonical taste-based discrimination has an intuitive representation in terms of rational expected utility maximization when Y_i^* is binary:

Lemma 3. If $Y_i^* \in \{0, 1\}$ then a judge can be said to be engaged in canonical taste-based discrimination if and only if there exists an S_i and $\mathcal{U}(y, d, r) : \{0, 1\}^2 \times \{w, b\} \to \mathbb{R}$ such that $\mathcal{U}(0, 1, r) - \mathcal{U}(0, 0, r) < 0$ for each r, $(\mathcal{U}(0, 1, r) - \mathcal{U}(0, 0, r)) \,/\, (\mathcal{U}(1, 1, r) - \mathcal{U}(0, 1, r))$ depends on r, and

$$D_i \in \arg\max_{D(\cdot) \in \mathbf{D}} E\left[\mathcal{U}(Y_i, D(R_i, S_i), R_i) \mid R_i, S_i\right], \tag{3}$$

where $\mathbf{D}$ is as in Lemma 2.

Here we consider an expanded set of utility functions $\mathcal{U}(\cdot)$ which depend explicitly on race. In particular, we consider those which yield a differential utility cost to treatment when $Y_i^* = 0$ (i.e. $\mathcal{U}(0, 1, r) - \mathcal{U}(0, 0, r)$), relative to the utility cost of the undesired outcome when $D_i = 1$ (i.e. $\mathcal{U}(1, 1, r) - \mathcal{U}(0, 1, r)$), by race r. The appendix proof shows that this condition can be interpreted as a racial difference in the relative utility cost of type-I error (failing to treat an individual with $Y_i^* = 0$) and type-II error (treating an individual with $Y_i^* = 1$). As with Lemma 2, the appendix shows that a similar result follows in the general case of non-binary Y_i^* whenever we restrict to utility functions that are linear in y.

Lemmas 2 and 3 make clear that a categorization of judge behavior by accurate statistical discrimination vs. taste-based discrimination is incomplete, due to the restriction of rational beliefs and utility functions with limited arguments. I thus define a third residual category:

Definition 4. The judge can be said to have *biased beliefs* if she is neither engaged in accurate statistical discrimination nor canonical taste-based discrimination.

While the classic theoretical analysis of Becker (1957), Phelps (1972), Arrow (1973), and Aigner and Cain (1977) only considered agents acting on rational beliefs, in practice judges may base their decisions on inaccurate beliefs or racial stereotypes (Bordalo et al., 2016). The notion of biased beliefs captures this form of racial discrimination, as well as other departures from accurate statistical or taste-based discrimination. For example, it could capture a judge that strays from her mandate and bases decisions on an accurate prediction of some other $\tilde{Y}_i \neq Y_i^*$. Kleinberg et al. (2017b) label this "omitted payoff bias."

Distinguishing between racial bias due to racial animus and biased beliefs is challenging, as preferences and beliefs can manifest equivalently in a judge's decisions (Bohren et al., 2020). To see this equivalence simply, consider a parametric decision-making model inspired by Aigner and Cain (1977), in which $Y_i^* \mid R_i$ is normally distributed and the judge observes both race and a noisy signal $S_i = Y_i^* + \eta_i$, with $\eta_i \mid R_i, Y_i^* \sim N(0, \sigma^2)$. In this case $E[Y_i^* \mid R_i, S_i] = (1 - \lambda_{R_i})\bar{\mu}_{R_i} + \lambda_{R_i} S_i$ where $\bar{\mu}_{R_i} = E[Y_i^* \mid R_i]$ and $\lambda_{R_i} \in (0, 1)$ is the slope coefficient from a regression of Y_i^* on S_i. A judge engaging in canonical taste-based discrimination will thus have a decision rule of $D_i = \mathbf{1}[(1-\lambda_{R_i})\bar{\mu}_{R_i} + \lambda_{R_i} S_i \leq t(R_i)]$ for some $t(w) \neq t(b)$. But an equivalent decision rule can be obtained from a judge with biased beliefs. For example a judge with a prior belief that $E[Y_i \mid R_i]$ in fact equals $\tilde{\mu}_{Ri} = \bar{\mu}_{R_i} - \frac{t(R_i)}{1-\lambda_{R_i}}$, but who observes the same signal and otherwise understands the statistical environment, will have an identical decision rule when attempting to statistically discriminate with a threshold of $t = 0$. Thus, it is impossible here to distinguish between a judge exhibiting racial bias because of racial tastes or inaccurate beliefs by her behavior alone: i.e., by her ranking U_i. It is further unclear at this stage whether these judges can be distinguished from a judge engaged in accurate statistical discrimination, and if so by what information. I next show how the marginal outcome test answers this question.

3.2 Interpreting Marginal Effects

I now present two results relating the marginal outcome test to the previous definitions.

Proposition 1. If the judge is engaged in canonical taste-based discrimination then $\beta \neq 0$.

Proposition 2. If $\beta \neq 0$ then the judge's decisions exhibit racial bias.

Proposition 1 shows that a judge who accurately predicts Y_i^* but places different thresholds on these predictions by race (as in Becker (1957)) will "fail" the outcome test, in that her marginal treatment effects will differ by race. Proposition 2 shows instead that a failure of the test implies a lack of accurate race-neutral decision-making (as in Phelps (1972), Arrow (1973), and Aigner and Cain (1977)). In other words, Proposition 2 shows that the marginal outcome test can reject the null hypothesis of accurate statistical discrimination, while Proposition 1 shows that the alternative hypothesis of canonical taste-based discrimination can drive such rejections. That Proposition 2 is not the strict converse of Proposition 1 reflects the fact that test rejections can also arise from biased beliefs.

The proofs of both propositions are simple, following just from the law of iterated expectations. For Proposition 1, we have under canonical taste-based discrimination

$$\begin{aligned} E[Y_i^* \mid U_i = 0, R_i] &= E[Y_i^* \mid E[Y_i^* \mid R_i, S_i] = t(R_i), R_i] \\ &= E[E[Y_i^* \mid R_i, S_i] \mid E[Y_i^* \mid R_i, S_i] = t(R_i), R_i] = t(R_i), \end{aligned} \tag{4}$$

showing that the marginal outcome test yields $\beta = t(w) - t(b) \neq 0$. In particular, canonical taste-based discrimination against Black (white) defendants manifests as $\beta > 0$ ($\beta < 0$). The proof to Proposition 2 similarly follows by showing that accurate statistical discrimination implies $\beta = 0$. Note that together Propositions 1 and 2 imply that the decisions of a judge engaged in canonical taste-based discrimination exhibit racial bias, as promised above.

In the binary Y_i^* case, Lemma 2 and Proposition 2 show that the marginal outcome test can reject that a judge's decisions are derived from her optimizing some expected race-neutral utility function of misconduct outcomes and treatment $\mathcal{U}(y, d)$, given some race-inclusive information set (R_i, S_i). That is, she is either acting rationally according to a race-specific utility function $\mathcal{U}(y, d, r)$ as in Lemma 3, which causes $\beta \neq 0$ by Proposition 1, or she has biased beliefs. As in the general case, biased beliefs could mean a failure to optimize her intended mandate of Y_i because of inaccurate predictions or because she predicting some other outcome that is not her mandate (i.e. omitted payoff bias). As the simple normal-signal example in Section 3.1 shows, it is generally difficult to distinguish between bias due to tastes and beliefs, and the marginal outcome test is correspondingly uninformative.

I next show that additional information on the race-specific marginal treatment effects frontiers allows a more nuanced analysis of racial bias:

Proposition 3. There exists an S_i such that $U_i = E[Y_i^* \mid R_i, S_i] - t(R_i)$ for some $\{t(w), t(b)\}$ if and only if $\mu(m, r)$ is strictly increasing in m for each $r \in \{w, b\}$.

This result shows that the slope of the MTE frontiers determine whether or not a judge's behavior can be rationalized by accurate beliefs (either statistical or taste-based discrimination). Intuitively, when the judge is acting on an accurate prediction of Y_i^*, perhaps with

Table 1: Possible Characterizations of Judge Behavior from Marginal Effects

		Marginal Outcome Test	
		Pass ($\beta = 0$)	Fail ($\beta \neq 0$)
MTE Slopes $\mu(m, r)$	Increasing in m	Accurate Stat. Disc.	Canonical T.-B. Disc.
	Non-Increasing	Biased Beliefs	Biased Beliefs

Notes: This table illustrates possible descriptions of judge decision-making, depending on whether or not the marginal outcome test passes (columns) and whether or not both race-specific MTE frontiers are strictly increasing (rows). Underlined descriptions indicate forms of racial bias.

race-specific thresholds, her race-specific MTEs must be increasing in the share of treated individuals. Proposition 3 shows that the converse also holds: when the race-specific MTE frontiers $\mu(m, r)$ are increasing in the treated share m, there exists an information set $\{R_i, S_i\}$ that can rationalize the judge as acting on accurate predictions of Y_i^*.[6] To my knowledge this converse is novel to the recent literature on interpreting such tests, though Marx (2018), Canay et al. (2020), and Gelbach (2021) prove related results on how canonical taste-based discrimination yields monotone MTE functions.[7]

Table 1 summarizes the possible characterizations of judge behavior from knowledge of both the slope of $\mu(m, r)$ and β. One consequence of Proposition 3 is that when $\mu(m, r)$ is increasing and the marginal outcome test "passes" (i.e. $\beta = 0$) we can say the judge is engaged in accurate statistical discrimination ($t(w) = t(b)$). A second consequence is that when $\mu(m, r)$ is increasing and the marginal outcome test "fails" (i.e. $\beta \neq 0$) we can say the judge is engaged in canonical taste-based discrimination ($t(w) \neq t(b)$). A non-increasing MTE frontier rules out these explanations, implying biased beliefs regardless of β. This could again either imply inaccurate predictions of Y_i^* or omitted payoff bias.

It is worth emphasizing that these conclusions of accurate statistical discrimination, canonical taste-based discrimination, and biased beliefs reflect what can be said about the judge's behavior from her actions, and not necessarily her "true" or intended behavior. For example, a judge who fails to accurately predict Y_i^* can still base her decisions on a ranking U_i that yields a strictly increasing MTE frontier because of some offsetting preferences among individuals of each race. If she equalizes marginal outcomes despite this combination of inaccurate prediction and omitted payoffs (i.e. $\beta = 0$) we would not be able to distinguish her from a judge who is accurately statistically discriminating absent further information.

[6]The requirement that $\mu(m, r)$ be strictly increasing in m reflects the requirement that $U_i \mid R_i$ be continuously distributed. It can be shown by a straightforward extension of Proposition 3 that weakly increasing $\mu(m, r)$ are equivalent to the existence of an S_i with $U_i = E[Y_i^* \mid R_i, S_i] - t(R_i) - \epsilon_i$ for some independent and continuously distributed ϵ_i that breaks indifferences in $E[Y_i^* \mid R_i, S_i] - t(R_i)$.

[7]In the setting of traffic stops, Marx (2018) shows that unconditional "hit rate" functions increase concavely with search probabilities under canonical taste-based discrimination. Canay et al. (2020) and Gelbach (2021) derive similar results and show how they imply monotone MTE functions.

Similarly, if her inaccurate ranking fails to equalize marginal outcomes ($\beta \neq 0$) we would not be able to say that she is not engaged in canonical taste-based discrimination. An "as-if" characterization of behavior is inherent to the exercise of inferring bias from behavior, since preferences and beliefs are typically difficult to disentangle. Recall again the simple normal-signal model of Section 3.1, where canonical taste-based discrimination and biased beliefs are indistinguishable except by knowing the judge's true intentions.[8]

Despite not knowing a judge's true intentions, the ability to rule out accurate statistical discrimination and give an "as-if" characterization of behavior can be valuable in practice. In the pretrial release setting, the finding of racial bias (by $\beta \neq 0$) may call for policy reforms that attempt to align judge behavior more closely with accurate statistical discrimination: for example, by providing more accurate algorithmic predictions of pretrial misconduct (e.g. Kleinberg et al., 2017b) or some form of implicit bias training (e.g. Morris, 2020). The finding of biased beliefs (by decreasing MTE frontiers) is perhaps even more consequential, as it suggests judges may be badly misinterpreting their mandate or mispredicting outcomes.

To summarize, we have seen a general interpretation of marginal outcome tests in terms of policy-relevant concepts linked to classic economic models of racial bias and statistical discrimination. I emphasize that this interpretation requires no assumptions on judge decision-making beyond the existence of some continuous ranking (Assumption 1). I next contrast this interpretation with the analysis of Canay et al. (2020), who propose a behavioral restriction necessary to interpret marginal outcome tests in terms of a different definition of bias. I further relate this section's definition with others in the recent literature.

4 Relationships to Other Analyses

This section first contrasts the previous analysis with the alternative definition of racial bias in Canay et al. (2020)—hereafter CMM—and the restriction on decision-making which they claim is necessary to restore the "validity" of marginal outcome tests. I then discuss connections to other definitions from the recent economics and computer science literatures.

4.1 Canay et al. (2020)

CMM consider a model of judge-decision making that can be written

$$D_i = \mathbf{1}[\Lambda(R_i, V_i) \leq \tau(R_i, V_i)], \tag{5}$$

[8]Bohren et al. (2020) discuss how biased beliefs and animus can be potentially distinguished through an experimental information provision (see also Bottan and Perez-Truglia (2020)). Such interventions may change the judge's ranking U_i, and are thus outside the scope of this paper's framework which treats the distribution of U_i as fixed.

where $\Lambda(R_i, V_i) = E[Y_i^* \mid R_i, V_i]$ is interpreted as an accurate prediction of Y_i^*, which CMM refer to as the "cost function." The corresponding "benefit function" $\tau(R_i, V_i)$ captures the residual determinants of the judge's decision, and is allowed to depend on both race R_i and what CMM refer to as "non-race characteristics" V_i.[9] In terms of the general model presented in Section 2, the CMM model is nested by $U_i = \Lambda(R_i, V_i) - \tau(R_i, V_i)$.

CMM use the benefit function to define a different notion of racial bias than in Definition 2. They state that the judge is biased against Black individuals if $\tau(b, v) < \tau(w, v)$ for all v in the support of V_i, is biased against white individuals if $\tau(b, v) > \tau(w, v)$ for all v in the support of V_i, and is racially unbiased if $\tau(w, v) = \tau(b, v)$ for all v in the support of V_i.[10] They then show that the marginal outcome test is generally uninformative for this definition of bias, except under a restriction on the benefit function which removes V_i: i.e.

$$D_i = \mathbf{1}[\Lambda(R_i, V_i) \leq \tau(R_i)], \tag{6}$$

which has the form of an extended Roy model (Heckman and Vytlacil, 2007; D'Haultfœuille and Maurel, 2013).[11] CMM thus conclude that in order to interpret the marginal outcome test in terms of racial bias, researchers must justify the extended Roy model as a behavioral restriction—a model which they view as very restrictive.

It is first worth noting that, in contrast with the Section 3 analysis, the CMM bias definition generally requires the information set $\{R_i, V_i\}$ to be specified. Many pairs of V_i and $\tau(r, v)$ functions will generally yield equivalent decision rules, so a definition of bias based on $\tau(r, v)$ is only unambiguous when V_i is defined. An arguable strength of the Section 3 definition, which instead considers the existence of any V_i rationalizing decisions by accurate statistical discrimination, is that such specification is unnecessary.[12]

To see why the specification of V_i matters for the CMM analysis, consider a simple example in which a judge engages in canonical taste-based discrimination, with her decisions given by $D_i = \mathbf{1}[E[Y_i^* \mid R_i, S_i] < t(R_i)]$ for some signal S_i and thresholds $t(w) > t(b)$. Clearly, if $V_i = S_i$, the CMM definition aligns with Definition 2 in saying that the judge is racially biased against Black individuals. But an identical set of decisions are generated by

[9] CMM index the benefit function by the identity of the judge (see their Equation (1)). I suppress this indexing since, as in CMM's conceptual analysis (their Section 3), I here consider a single judge.

[10] Unlike the Section 3 typology this definition leaves an omitted category, of judges where $\tau(w, v)$ and $\tau(b, v)$ are neither always higher than, lower than, or equal to each other.

[11] The extended Roy model is not the only restriction that CMM consider to restore "logical validity" of the marginal outcome test. However they note that the alternative restriction (that the cost function does not depend on race) is likely harder to justify in practice, since race is likely to predict Y_i^*.

[12] Another strength of the Section 3 analysis is that it allows for evaluations of bias along multiple lines: i.e., a marginal outcome test that splits the sample by race can be interpreted alongside a test that splits by an individual's sex. In contrast the extended Roy model restriction of CMM cannot hold for both characteristics jointly when sex is included in V_i, making at least one of these tests uninformative by their analysis.

a model in which the judge has a broader information set of $V_i = \{S_i, T_i\}$ for some other signal T_i and sets $\tau(R_i, V_i)$ appropriately: i.e.,

$$D_i = \mathbf{1}[\underbrace{E[Y_i^* \mid R_i, V_i]}_{\equiv \Lambda(R_i, V_i)} < \underbrace{t(R_i) + E[Y_i^* \mid R_i, V_i] - E[Y_i^* \mid R_i, S_i]}_{\equiv \tau(R_i, V_i)}]. \tag{7}$$

With this rationalization, of the same decisions, we can no longer say the judge is biased under CMM's definition, since $\tau(r, v)$ need not always be larger for white (or Black) individuals.

Specifying V_i is also important for CMM's claim on the restrictiveness of the extended Roy model. Indeed, the above Proposition 3 shows that whenever the race-specific MTE curves are strictly increasing there exists a V_i such that decisions have an extended Roy model representation. Thus, when V_i is unspecified, the CMM restriction can be imposed without loss to the decision-making process in this case. Our two definitions of bias, and two conclusions on the marginal outcome tests' ability to detect bias, would then agree.

While taking at least a conceptual stand on the judge's information set V_i is inherent to the CMM analysis, it may be challenging or undesirable. The challenge comes from the fact that in practice decision-makers may combine a wide range of objective and subjective information that can be difficult to specify in V_i, even conceptually. In the pretrial example, one could imagine collecting all objective (not necessarily observed) measures of defendant and case characteristics—such as demographics, charge type, and previous criminal activity—in V_i. This fixes the $\tau(r, v)$ functions, and thus the CMM bias definition, as residuals from the objective prediction $\Lambda(R_i, V_i) = E[Y_i^* \mid R_i, V_i]$, in essence benchmarking the judge's behavior to an objective (perhaps unmeasured) risk score. But in practice pretrial judges may also base decisions on face-to-face interviews with defendants, and vary in what "soft" information or signals are taken from such interactions. It can be difficult to conceptualize a well-defined notion of bias, from a well-defined V_i, in such cases.

Moreover, even when it is possible to conceptualize collecting all available information in V_i, a bias definition based on this set may be unsatisfying when judges vary in their skill at incorporating signals. A given pretrial judge may, for example, be highly skilled at inferring misconduct potential from a defendant's demeanor, while another judge chooses to rely more on the "hard" information in a defendant's criminal record.[13] To see how such variation complicates the CMM analysis, suppose we are able to put (along with all defendant and case characteristics) a "complete" record of objective information from pretrial interviews in V_i, including sophisticated data on, e.g., a defendant's demeanor and ability to answer standard questions. But suppose a judge is only partially attentive to this rich data, observing a signal S_i which is strictly less informative: i.e., $V_i = \{S_i, T_i\}$, where T_i is the subset of information

[13]Evidence for such variation in predictive skill has recently been documented in the pretrial setting (Arnold et al., 2020a) and in medical testing decisions (Chan et al., 2020).

for individual i that is missed by the judge (such as the answers to questions she forgets to ask, or elements of the defendant's demeanor she overlooks). Suppose the judge uses this signal to engage in accurate statistical discrimination, forming an accurate but inefficient posterior of $E[Y_i^* \mid R_i, S_i] \neq E[Y_i^* \mid R_i, V_i]$. Her otherwise race-neutral decision rule is then

$$D_i = \mathbf{1}[E[Y_i^* \mid R_i, S_i] \leq t] = \mathbf{1}[\underbrace{E[Y_i^* \mid R_i, V_i]}_{\equiv \Lambda(R_i, V_i)} \leq \underbrace{t + E[Y_i^* \mid R_i, V_i] - E[Y_i^* \mid R_i, S_i]}_{\equiv \tau(R_i, V_i)}], \quad (8)$$

with the final expression having the form of CMM's representation (5). This decision-rule generally does not satisfy the extended Roy model restriction, since $\tau(R_i, V_i)$ will generally vary nontrivially with V_i. The CMM analysis will thus conclude the marginal outcome test is uninformative of the judge's bias, even though she is engaged in classic statistical discrimination. In contrast, the Section 3 analysis would say the marginal outcome test "passes" in this scenario, with $\beta = 0$ indicating a lack of racial bias. This conclusion appears warranted given the judge's accurate beliefs and race-neutral threshold t.

In summary, there are several important differences between the Section 3 analysis of marginal outcome tests and CMM's. The former does not require specification of the judge's information set for bias to be well-defined, and for β to detect canonical taste-based discrimination (in the sense of Becker (1957)). At the same time, the Section 3 interpretation can "clear" an agent engaged in classic statistical discrimination (in the sense of Aigner and Cain (1977)) from suspicions of bias, even when her skill at using an objective set of available information is imperfect. Neither of these conclusions require the extended Roy model restriction of CMM, which we have shown is without loss to impose when the judge's information is unspecified and the race-specific MTE curves are increasing.

4.2 Unwarranted Disparity

One critique of marginal outcome tests, given the Section 3 analysis, is that they capture a narrow form of racial discrimination. In particular, the kinds of accurate statistical discrimination which cause such tests to "pass" can be viewed as undesirable or even illegal in many settings, including potentially the pretrial setting (Yang and Dobbie, 2020). Motivated by this shortcoming, Arnold et al. (2020a) propose an alternative measure of discrimination which aligns with Aigner and Cain (1977) in capturing any "unwarranted" racial disparities in treatment rates that cannot be justified by differences in an individual's "qualification" for treatment (as given by their Y_i^*). Formally, they define unwarranted disparity as

$$\Delta = E[E[D_i \mid R_i = w, Y_i^*] - E[D_i \mid R_i = b, Y_i^*]], \quad (9)$$

or the average treatment rate disparity of white and Black individuals with identical Y_i^*. Arnold et al. (2020a) show how this notion of discrimination can capture racial bias, defined similarly to the Section 3 analysis, as well as accurate statistical discrimination.[14]

Because it is based on a narrower measure of discrimination, a finding of no racial bias (i.e. $\beta = 0$) need not imply the equal treatment of equally qualified individuals (i.e. $\Delta = 0$). Interestingly, however, the converse also holds: a finding of no unwarranted disparity need not imply a lack of racial bias. This follows from a general trade-off between such "fairness" notions and unbiased prediction in this kind of decision-making problem (Kleinberg et al., 2017a). Some bias "at the margin" of treatment (as captured by β) is generally required for fair treatment "on average" (as captured by Δ).

Feigenberg and Miller (2020) discuss a scenario in which the difference between a measure like β and a measure like Δ is especially stark. In an analysis of Texas traffic stop data, they find that the marginal outcome test passes despite significant unwarranted disparities in state trooper stop rates between white and Black drivers. The reason for these disparities appears to be that, conditional on observables, the typical trooper effectively searches drivers at random but with minority drivers searched more. Here it is worth noting that while marginal outcome tests cannot generally detect such unequal treatment, knowledge of the race-specific MTE frontiers can: under random decision-making the $\mu(m, r)$ will be constant over m and equal to the average treatment effect $E[Y_i^* \mid R_i = r]$.

4.3 Algorithmic Calibration

While the Section 3 analysis presumed a human decision-maker, its definition of bias has a close link to the notion of "calibration" in the literature on algorithmic decision-making (Barocas and Selbst, 2016; Kleinberg et al., 2017a; Chouldechova, 2017). This literature typically considers decisions based on some statistical prediction of Y_i^*, $P(X_i)$, from some observed covariates X_i. For example, $P(X_i)$ may be generated by a machine learning algorithm. Calibration bias is then defined by a racial disparity in prediction error,

$$\Gamma(p) = E[Y_i^* \mid P(X_i) = p, R_i = w] - E[Y_i^* \mid P(X_i) = p, R_i = b], \tag{10}$$

for some p (Obermeyer et al., 2019). It is immediate that this definition coincides with the Section 3 definition of racial bias when treatment decisions are algorithmic: formally, if $D_i = \mathbf{1}[P(X_i) \leq t]$ for some t then $\beta = \Gamma(t)$. A marginal outcome test applied to such an algorithmic decision-rule therefore reveals calibration bias.

An important feature of investigations of algorithmic bias is that while the ranking $P(X_i)$

[14] This Δ is also closely related to notions of algorithmic discrimination in the computer science literature, in particular "conditional procedure accuracy equality" (Berk et al., 2018).

is known, and may even be under the researcher's control, it need not coincide with an accurate prediction $E[Y_i^* \mid X_i]$. This is because the treatment effects Y_i^* need not be observed for some or all individuals, a fundamental challenge that here Lakkaraju et al. (2017) term the "selective labels problem." This feature highlights the need for allowing for what Section 3 calls biased "beliefs" in testing for calibration bias, although this terminology is less natural for algorithmic decisions. For especially severe cases of the selective labels problem, calibration bias can manifest with non-increasing race-specific MTE frontiers.[15]

5 Identification Strategies

I lastly turn to the question of how marginal outcome tests can be conducted in practice. Lemma 2 shows that identification of β follows from the identification of judge- and race-specific MTE curves, at least near the judge's treatment cutoff. The recent empirical literature has devised several strategies leveraging different sources of variation. In some cases, these strategies permit full identification of the MTE frontiers and thus the more nuanced analysis of behavior summarized in Table 1.

It is first worth considering an experimental ideal, in which a judge is induced to increase or decrease her treatment rate among a random set of both white and Black individuals. By comparing the "treatment" and "control" groups within each race, one may obtain an experimental estimate of race-specific treatment effects near the judge's cutoff. For small changes these effects can be interpreted as the two marginal treatment effects that enter β: the finding of a larger effect among white individuals relative to Black individuals would be consistent, for example, with canonical taste-based discrimination in inaccurate stereotypes favoring white individuals, while the failure to reject $\beta = 0$ would be consistent with accurate statistical discrimination. Larger changes may require parametric assumptions to interpret average effects in terms of marginal effects, but may trace out larger portions of the MTE curves that can be used to diagnose the extent of biased beliefs. Multiple treatment groups could be contrasted with a single control group within each race to fit polynomial race-specific MTE curves, for example. Per Proposition 3, one cannot rule out that the judge is acting on accurate predictions of Y_i^* when the estimated curves are increasing.

In practice, where the scope for such experimentation is limited, researchers may leverage quasi-experimental or observational variation in judge- and race-specific treatment rates. Feigenberg and Miller (2020), for example, use the variation in race-specific search rates for a given state trooper to estimate within-trooper MTE frontiers and study issues related

[15] Kleinberg et al. (2017b) and Arnold et al. (2021) discuss quasi-experimental solutions to the selective labels problem in the pretrial setting. The former show how algorithms can be used to improve human decision-making in spite of this problem, while the latter show how the problem can be overcome to estimate measures like Δ and $\Gamma(p)$.

to racial bias. They show that this variation is largely idiosyncratic with respect to motorist characteristics, as would be guaranteed in an idealized experimental manipulation of trooper search rates. Grau and Vergara (2021) instead propose an observational approach to identifying marginal outcomes, which can in principle be conducted for an individual judge.

When it is difficult to justify using within-judge treatment rate variation, a natural quasi-experimental alternative is to leverage exogenous variation across judges. Arnold et al. (2018) introduce this approach in the pretrial setting, where judges are plausibly as-good-as-randomly assigned to cases after adjusting for courtroom and time fixed effects. The key assumption they invoke to leverage this across-judge variation is conventional MTE monotonicity, which restricts the race-specific MTE curves to be common across judges. Under this assumption the different as-good-as-randomly assigned judges trace out a single race-specific MTE frontier, which can be used to both detect racial bias and probe the extent of biased beliefs.[16] A conceptually similar approach is taken by Marx (2018) in the setting of police search. Feigenberg and Miller (2020) show that an analysis leveraging the quasi-random assignment of state troopers yields similar estimates as their within-trooper analysis, building confidence in the common MTE assumption in their setting.

In some settings, however, the assumption of common MTE frontiers may be untenable. In terms of the behavioral models in Section 3, this assumption generally restricts different judges to act on a common prediction of Y_i^*, given common signals S_i. In reality judges are likely to vary in their predictive skill, inducing different judge-specific MTE frontiers that violate the monotonicity assumption used by Arnold et al. (2018) and Marx (2018).[17]

A solution to this challenge is given by Arnold et al. (2020a), who develop a hierarchical MTE model to estimate racial discrimination and bias from the as-good-as-random assignment of pretrial judges. The model specifies race-specific distributions of signals, and thus distributions of MTE curves, across judges, in violation of conventional MTE monotonicity; Arnold et al. (2020a) show that this distribution can be estimated from certain reduced-form moments. Notably, in contrast to the conventional MTE approach of Arnold et al. (2018), the Arnold et al. (2020a) approach uses parametric restrictions. This reflects the general fact that other assumptions are required once the uniformity condition of monotonicity is relaxed.[18] In their setting Arnold et al. (2020a) show that uniformity is an especially unap-

[16] In addition to monotonicity, Arnold et al. (2018) discuss assumptions that are needed to overcome the core inframarginality problem of only observing a finite number of judges with discrete differences in race-specific treatment rates. I note that this problem exists in *any* evaluation of marginal effects from discrete treatment rate interventions, including the experimental ideal.

[17] Marx (2018) considers a model where judges can act on different but equally informative signals. This yields a common MTE frontier despite a violation of strict monotonicity (where judges observe identical signals). See Chan et al. (2020) and Feigenberg and Miller (2020) for discussions of similar assumptions.

[18] While Arnold et al. (2020a) specify their model in terms of accurate statistical discrimination and taste-based discrimination, they note that its parameterization also allows judges to have biased beliefs (similar to the simple normal-signal example in Section 3). In practice they find upward-sloping MTE frontiers.

pealing restriction, as there appears to be sizable variation in judge predictive skill.

Overall, the recent empirical literature leveraging quasi-experimental variation to conduct marginal outcome tests (and related exercises) is nascent and only represents some of the possible identification strategies. Ways to relax the parametric structure of Arnold et al. (2020a) and estimate racial bias from as-good-as-random judge assignment without a classic MTE monotonicity assumption seem worth future study. Settings where within-judge variation can be credibly used to avoid such across-judge restrictions also appear valuable.

6 Conclusion

Since Becker (1957), economists have long theorized about the ways in which racial biases can differentially affect the outcomes of white and Black individuals at the margin of treatment. A recent empirical literature has proposed using quasi-experimental variation to estimate such effects, drawing on the tractable MTE framework and avoiding the specification of complete decision-making models. It is not always clear, however, what this approach generally reveals about racial bias.

This paper has presented a formal definition of racial bias which can be revealed by the marginal outcome test: the lack of accurate statistical discrimination, which Phelps (1972), Arrow (1973), and Aigner and Cain (1977) derived to contrast with Becker's theory of taste-based discrimination. We have seen how this definition encompasses different forms of bias, from canonical taste-based discrimination to inaccurate beliefs or stereotypes, and how in some cases these explanations can be distinguished by knowing complete MTE frontiers. This analysis contrasts with the alternative bias definition in Canay et al. (2020), which leads to a more pessimistic view of marginal outcome tests. The notion of bias which I show marginal outcome tests reveal also contrasts with other recent definitions of discrimination, while aligning with a notion of algorithmic bias from the computer science literature. Finally, we have seen some ways that marginal outcome tests have been conducted in the recent empirical literature, under different identifying assumptions. More work is needed on developing tractable restrictions on judge MTE curves which can reveal racial bias while also allowing for heterogeneity in judge predictive skill.

References

Aigner, D. J. and G. G. Cain (1977): "Statistical Theories of Discrimination in Labor Markets," *Industrial and Labor Relations Review*, 30, 175–187.

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

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

Arnold, D., W. Dobbie, and P. Hull (2020a): "Measuring Racial Discrimination in Bail Decisions," *NBER Working Paper No. 26999*.

——— (2021): "Measuring Racial Discrimination in Algorithms," *American Economic Association: Papers & Proceedings*.

Arnold, D., W. Dobbie, and C. S. Yang (2018): "Racial Bias in Bail Decisions," *Quarterly Journal of Economics*, 133, 1885–1932.

——— (2020b): "Comment on Canay, Mogstad, and Mountjoy (2020)," *Mimeo*.

Arrow, K. J. (1973): "The Theory of Discrimination," in *Discrimination in Labor Markets*, ed. by O. Ashenfelter and A. Rees, Princeton University Press, 3–33.

Barocas, S. and A. D. Selbst (2016): "Big Data's Disparate Impact," *California Law Review*, 104, 671.

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

Berk, R., H. Heidari, S. Jabbari, M. Kearns, and A. Roth (2018): "Fairness in Criminal Justice Risk Assessments: The State of the Art," *Sociological Methods & Research*, 1–42.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2020): "Inaccurate Statistical Discrimination: An Identification Problem," *NBER Working Paper No. 25935*.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016): "Stereotypes," *The Quarterly Journal of Economics*, 131, 1753–1794.

Bottan, N. L. and R. Perez-Truglia (2020): "Betting on the House: Subjective Expectations and Market Choices," *NBER Working Paper No. 27412*.

Canay, I. A., M. Mogstad, and J. Mountjoy (2020): "On the Use of Outcome Tests for Detecting Bias in Decision Making," *NBER Working Paper No. 27802*.

Chan, D., M. Gentzkow, and C. Yu (2020): "Selection with Variation in Diagnostic Skill: Evidence from Radiologists," *NBER Working Paper No. 26467*.

Chouldechova, A. (2017): "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *Big Data*, 5, 153–163.

D'Haultfœuille, X. and A. Maurel (2013): "Inference on an Extended Roy Model, with an Application to Schooling Decisions in France," *Journal of Econometrics*, 174, 95–106.

Feigenberg, B. and C. Miller (2020): "Racial Disparities in Motor Vehicle Searches Cannot Be Justified By Efficiency," *Unpublished Working Paper*.

Gelbach, J. B. (2021): "Testing Economic Models of Discrimination in Criminal Justice," *Unpublished Working Paper*.

Grau, N. and D. Vergara (2021): "An Observational Implementation of the Outcome Test with an Application to Ethnic Prejudice in Pretrial Detentions," *Unpublished Working Paper*.

Heckman, J. J. and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738.

——— (2007): "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators, to Evaluate Social Programs, and to Forecast their Effects in New Environments," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. Leamer, Elsevier, 4875–5143.

Kleinberg, J., 20020503, S. Mullainathan, and M. Raghavan (2017a): "Inherent Trade-Offs in Algorithmic Fairness," *Proceedings of Innovations in Theoretical Computer Science*, 43:1–43:23.

Kleinberg, J., 20020603, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2017b): "Human Decisions and Machine Predictions," *The Quarterly Journal of Economics*, 133, 237–293.

Knowles, J., N. Persico, and P. Todd (2001): "Racial Bias in Motor Vehicle Searches: Theory and Evidence," *Journal of Political Economy*, 109, 203–229.

Lakkaraju, H., J. Kleinberg, J. Leskovec, J. Ludwig, and S. Mullainathan (2017): "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables," *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 275–284.

Marx, P. (2018): "An Absolute Test of Racial Prejudice," *Unpublished Working Paper*.

Morris, A. (2020): "Confronting Implicit Bias: Texas Judges Would Get Annual Training Under New Resolution," *Texas Lawyer*, `https://bit.ly/36REEhI`.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan (2019): "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science*, 366, 447–453.

Persico, N. and P. Todd (2006): "Generalizing the Hit Rate Tests for Racial Bias in Law Enforcement, with an Application to Vehicle Searches in Wichita," *Economic Journal*, 116, F351–F367.

Phelps, E. S. (1972): "The Statistical Theory of Racism and Sexism," *American Economic Review*, 62, 659–661.

Yang, C. and W. Dobbie (2020): "Equal Protection Under Algorithms: A New Statistical and Legal Framework," *Michigan Law Review*, 119, 291–396.

A Appendix Proofs

A.1 Lemma 1

Let $F_{U|R}(u, r)$ denote the cumulative distribution function for $U_i \mid (R_i = r)$. Then

$$D_i = \mathbf{1}[F_{U|R}(U_i, R_i) \leq F_{U|R}(0, R_i)] \equiv \mathbf{1}[W_i \leq F_{U|R}(0, R_i)] \tag{A1}$$

where $W_i \mid R_i \sim \mathrm{U}(0, 1)$. This defines the MTE frontiers $\mu(m, r)$. Furthermore,

$$F_{U|R}(0, r) = Pr(D_i = 1 \mid R_i = r) = \pi_r, \tag{A2}$$

so $E[Y_i^* \mid U_i = 0, R_i = r] = E[Y_i^* \mid W_i = \pi_r, R_i = r] = \mu(\pi_r, r)$. □

A.2 Lemma 2

First consider an arbitrary $\mathcal{U}(\cdot)$ and S_i. In the binary case, where $Y_i = D_i Y_i^*$, we have for any decision rule D_i,

$$\begin{aligned} \mathcal{U}(Y_i, D_i) &= \mathcal{U}(1,1) Y_i^* D_i + \mathcal{U}(0,1)(1 - Y_i^*) D_i + \mathcal{U}(0,0)(1 - D_i) \\ &= \alpha Y_i^* D_i - \gamma D_i + \kappa, \end{aligned} \tag{A3}$$

where $\alpha = \mathcal{U}(1,1) - \mathcal{U}(0,1) < 0$, $\gamma = \mathcal{U}(0,0) - \mathcal{U}(0,1)$, and $\kappa = \mathcal{U}(0,0)$. Therefore the optimization in Equation (1) can be written

$$\max_{D(\cdot) \in \mathbf{D}} \alpha D(R_i, S_i) E[Y_i^* \mid R_i, S_i] - \gamma D(R_i, S_i) + \kappa, \tag{A4}$$

It is clear that solutions to this problem are given by

$$\begin{aligned} D_i &= \mathbf{1}[\alpha E[Y_i^* \mid R_i, S_i] \geq -\gamma] \\ &= \mathbf{1}[E[Y_i^* \mid R_i, S_i] \leq t], \end{aligned} \tag{A5}$$

for $t = \gamma / \alpha$. Thus, if a judge's decisions can be written as in Equation (1) for some $\mathcal{U}(\cdot)$ and S_i, she can be said to be engaged in accurate statistical discrimination. Conversely, if a judge's decisions can be written as in Equation (A5) for some S_i and t, there exist utility functions $\mathcal{U}(\cdot)$ such that her decisions can be represented by Equation (1). □

Note that this equivalence means the maintained Assumption 1 implicitly restricts the set of $\mathcal{U}(\cdot)$ considered in Lemma 2. For example, for the judge's decisions to be non-degenerate we require $(\mathcal{U}(0,0) - \mathcal{U}(0,1))/(\mathcal{U}(1,1) - \mathcal{U}(0,1)) \in (0, 1)$. Note also that the same steps can

be used to show the equivalence between accurate statistical discrimination and solutions to Equation (1) in the general case of non-binary Y_i, when restricting to $\mathcal{U}(y,d)$ that are linear in y. Specifically, if $\mathcal{U}(y,d) = \alpha y + \mathcal{U}_0(d)$ for < 0 then

$$\mathcal{U}(Y_i, D_i) = \alpha Y_{i1} D_i + \mathcal{U}_0(1) D_i + \alpha Y_{i0}(1 - D_i) + \mathcal{U}_0(0)(1 - D_i) = \alpha Y_i^* D_i - \gamma D_i + \kappa, \quad \text{(A6)}$$

as before, except now $\gamma = \mathcal{U}_0(0) - \mathcal{U}_0(1)$ and $\kappa = \alpha Y_{i0} + \mathcal{U}_0(0)$.

A.3 Lemma 3

The proof follows similarly to that of Lemma 2. For arbitrary $\mathcal{U}(\cdot)$ and any decision rule D_i

$$\begin{aligned} \mathcal{U}(Y_i, D_i, R_i) &= \mathcal{U}(1,1,R_i) Y_i^* D_i + \mathcal{U}(0,1,R_i)(1 - Y_i^*) D_i + \mathcal{U}(0,0,R_i)(1 - D_i) \\ &= \alpha(R_i) Y_i^* D_i - \gamma(R_i) D_i + \kappa(R_i), \end{aligned} \quad \text{(A7)}$$

where $\alpha(R_i) = \mathcal{U}(1,1,R_i) - \mathcal{U}(0,1,R_i)$, $\gamma(R_i) = \mathcal{U}(0,0,R_i) - \mathcal{U}(0,1,R_i)$, and $\kappa(R_i) = \mathcal{U}(0,0,R_i)$. Therefore the optimization in Equation (3) can be written

$$\max_{D(\cdot) \in \mathbf{D}} \alpha(R_i) D(R_i, S_i) E[Y_i^* \mid R_i, S_i] - \gamma(R_i) D(R_i, S_i) + \kappa(R_i), \quad \text{(A8)}$$

It is clear that solutions to this problem are given by

$$D_i = \mathbf{1}[E[Y_i^* \mid R_i, S_i] \leq t(R_i)], \quad \text{(A9)}$$

where $t(R_i) = \gamma(R_i)/\alpha(R_i)$. Thus, if a judge's decisions can be written as in Equation (3) for some $\mathcal{U}(\cdot)$ and S_i with $\gamma(w)/\alpha(w) \neq \gamma(b)/\alpha(b)$, she can be said to be engaged in canonical taste-based discrimination. Conversely, if a judge's decisions can be written as in (A9) for some S_i and $t(w)(b)$, there exist utility functions $\mathcal{U}(\cdot)$ with $\gamma(w)/\alpha(w) \neq \gamma(b)/\alpha(b)$ such that her decisions can be represented by Equation (3). □

As in the proof to Lemma 2, this equivalence means the maintained Assumption 1 implicitly restricts the set of $\mathcal{U}(\cdot)$ considered in Lemma 3. Also that the same steps can be used to show the equivalence between accurate statistical discrimination and solutions to Equation (1) in the general case of non-binary Y_i, when restricting to linear $\mathcal{U}(y,d,r) = \alpha(r)y + \mathcal{U}_0(d,r)$.

Finally, I show that the condition of $\gamma(w)/\alpha(w) \neq \gamma(b)/\alpha(b)$ has an interpretation in terms of preferences over type-I and type-II error rates in the binary Y_i^* case. Note that without loss we can normalize one of the utility levels from the four combinations of (Y_i, D_i) to zero within each race. By normalizing $\mathcal{U}(0,1,r) = 0$ it can be seen that $\gamma(r)/\alpha(r)$ gives the ratio of utility from failing to treat an individual without the adverse outcome, $\mathcal{U}(0,0,r)$,

(i.e. type-I error) to utility from treating an individual with the adverse outcome, $\mathcal{U}(1,1,r)$, (i.e. type-II error).

A.4 Proposition 1

A full proof is given in the main text: see Equation (4).

A.5 Proposition 2

The proof is by contradiction. Suppose an S_i exists such that $U_i = E[Y_i^* \mid R_i, S_i] - t$ for some t. Then following Equation (4) we have by the law of iterated expectations

$$\begin{aligned} E[Y_i^* \mid U_i = 0, R_i] &= E[Y_i^* \mid E[Y_i^* \mid R_i, S_i] = t, R_i]] \\ &= E[E[Y_i^* \mid R_i, S_i] \mid E[Y_i^* \mid R_i, S_i] = t, R_i]] \\ &= t, \end{aligned} \tag{A10}$$

so $\beta = E[Y_i^* \mid U_i = 0, R_i = w] - E[Y_i^* \mid U_i = 0, R_i = b] = t - t = 0$. □

A.6 Proposition 3

First suppose there exists an S_i such that $U_i = E[Y_i^* \mid R_i, S_i] - t(R_i)$ for some $t(r)$. Then

$$\mu(m,r) = E[Y_i^* \mid E[Y_i^* \mid R_i = r, S_i] - t(r) = F_{U|R}^{-1}(m,r), R_i = r] \tag{A11}$$

where $F_{U|R}^{-1}(\cdot)$ denotes the inverse of the distribution function defined in the proof to Lemma 1. By the law of iterated expectations, as in Equation (4),

$$\mu(m,r) = F_{U|R}^{-1}(m,r) + t(r) \tag{A12}$$

which is strictly increasing in m, with $U_i \mid R_i$ continuously distributed.

Now suppose $\mu(m,r)$ is strictly increasing in m. Let $S_i = W_i$ be the uniformly distributed index in the MTE representation of the judge's decisions. Then $E[Y_i^* \mid R_i, W_i] = \mu(W_i, R_i)$ is itself a U_i which is continuously distributed given R_i. Furthermore

$$D_i = \mathbf{1}[W_i \leq \pi_{R_i}] = \mathbf{1}[E[Y_i^* \mid R_i, W_i] \leq \mu(\pi_{R_i}, R_i)] = \mathbf{1}[E[Y_i^* \mid R_i, S_i] \leq t(R_i)].$$

The first equality applies the strictly monotone $\mu(\cdot, R_i)$ to both sides of the inequality, while the second equality substitutes $W_i = S_i$ and defines $t(R_i) = \mu(\pi_{R_i}, R_i)$. □