NBER WORKING PAPER SERIES

# ON THE USE OF OUTCOME TESTS FOR DETECTING BIAS IN DECISION MAKING

Ivan A. Canay
Magne Mogstad
Jack Mountjoy

Working Paper 27802
http://www.nber.org/papers/w27802

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2020

On the Use of Outcome Tests for Detecting Bias in Decision Making
Ivan A. Canay, Magne Mogstad, and Jack Mountjoy
NBER Working Paper No. 27802
September 2020
JEL No. C21,C26,C51,J15,J16,J71,K14,K42


## ABSTRACT

The decisions of judges, lenders, journal editors, and other gatekeepers often lead to disparities in outcomes across affected groups. An important question is whether, and to what extent, these group-level disparities are driven by relevant differences in underlying individual characteristics, or by biased decision makers. Becker (1957) proposed an outcome test for bias leading to a large body of related empirical work, with recent innovations in settings where decision makers are exogenously assigned to cases and vary progressively in their decision tendencies. We carefully examine what can be learned about bias in decision making in such settings. Our results call into question recent conclusions about racial bias among bail judges, and, more broadly, yield four lessons for researchers considering the use of outcome tests of bias. First, the so-called generalized Roy model, which is a workhorse of applied economics, does not deliver a logically valid outcome test without further restrictions, since it does not require an unbiased decision maker to equalize marginal outcomes across groups. Second, the more restrictive "extended" Roy model, which isolates potential outcomes as the sole admissible source of analyst-unobserved variation driving decisions, delivers both a logically valid and econometrically viable outcome test. Third, this extended Roy model places strong restrictions on behavior and the data generating process, so detailed institutional knowledge is essential for justifying such restrictions. Finally, because the extended Roy model imposes restrictions beyond those required to identify marginal outcomes across groups, it has testable implications that may help assess its suitability across empirical settings.



Ivan A. Canay
Department of Economics
Northwestern University
2211 Campus Drive
Evanston, IL 60208
iacanay@northwestern.edu

Magne Mogstad
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637
and NBER
magne.mogstad@gmail.com

Jack Mountjoy
Booth School of Business
University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
jack.mountjoy@chicagobooth.edu

# 1 Introduction

The decisions of judges, police officers, teachers, doctors, lenders, landlords, journal editors, admissions committees, and other gatekeepers often lead to disparities in outcomes across affected groups. An important question is whether, and to what extent, these group-level disparities are driven by relevant differences in underlying individual characteristics, or by biased decision makers employing favoritism, animus, or inaccurate stereotypes that unfairly privilege particular groups at the expense of others.

To answer this question, it is necessary to first define what it means for a decision to be unbiased. This requires specifying what unbiased decision makers in a particular setting are supposed to be optimizing, what constraints they face, and what they should know at the time they make their decisions. Once this is specified, the analyst can derive optimality conditions for the decision maker's problem, and then attempt to check whether these conditions are consistent with data, separately for different groups affected by the decision. If these checks suggest that an unbiased decision maker could do better by changing how they treat members of a particular group, the analyst may conclude that this group is subject to bias.

This idea forms the basis of the outcome test for bias proposed by Becker (1957). Motivated by his work, numerous empirical studies have analyzed group-level disparities across a wide range of settings with the aim of learning about bias in decision making.[1] Much of this literature, however, does not actually specify the optimization problems faced by the relevant decision makers, which leaves the analyst and the reader without a clear framework for inferring (un)biased behavior from a given set of empirical results. Studies that do derive optimality conditions for unbiased behavior, moreover, usually face additional challenges when taking these conditions to data, including selection bias and, more subtly, inframarginality bias, whereby comparisons of group averages need not be informative about the marginal conditions that distinguish biased from unbiased behavior (Heckman and Siegelman, 1993; Heckman, 1998).

A handful of recent studies have made progress in addressing some of these challenges by focusing on outcome disparities in settings where decision makers are exogenously assigned to cases and vary progressively in their decision tendencies. Arnold, Dobbie, and Yang (2018) (hereafter

[1] See recent reviews by Bertrand and Duflo (2017) and Lang and Spitzer (2020).

ADY) exemplify this approach in their study of racial disparities in pre-trial bail decisions. Bail judges in their setting are as good as randomly assigned to defendants, meaning that any outcome disparities across decision makers can be attributed to the decision makers themselves and not the types of cases they handle. Importantly, ADY also make use of nearly continuous variation in release tendencies across judges, which helps address the inframarginality problem by isolating marginally released defendants of each race who would have been detained had they been assigned to a slightly less lenient judge. These advantageous institutional features are not confined to an obscure setting: nearly half a million defendants in the U.S. currently await trial from a jail cell, prior to any conviction, incurring large costs to taxpayers and to defendant livelihoods (Leslie and Pope, 2017; Dobbie et al., 2018).

In this paper, we carefully examine what can be learned about bias in decision making in this setting, and by doing so, offer a methodological blueprint for researchers considering the use of outcome tests to detect bias across a wide range of empirical environments. For concreteness, we conduct most of our analysis in the pre-trial release framework of ADY, as it offers an illuminating case study of the challenges that arise, even after important innovations, in inferring bias from outcome tests. In many other settings, the decision maker's objective is more complex, and problems of selection and inframarginality may arise, which will only exacerbate the difficulties we highlight here.

As detailed in Section 2, ADY write down a generalized Roy model of decision making in which bail judges weigh the expected benefits of releasing a given defendant against the expected cost, specified as the probability that the defendant commits any pre-trial misconduct. Expected costs and benefits of release vary across defendants with different characteristics, with race observed by both the judge and the analyst but some non-race characteristics observed only by the judge and not the analyst. ADY argue that such a model delivers an outcome test of racial bias. The claim is that bias can be inferred from the subsequent misconduct rates of marginally released black and white defendants, under the Beckerian intuition that judges biased against blacks will hold black defendants to a stricter standard and only release those with lower misconduct propensity relative to similar white defendants.

In Section 3, we show that this generalized Roy model of judge decision making fails to deliver the proposed outcome test. Given ADY's model, the test is logically invalid in both directions: it

may lead the analyst to infer bias when judges are actually unbiased, and infer no bias when judges are actually biased. The test may even indicate bias against one race when judges are actually biased against another. These logical invalidations arise because the specified decision model is not sufficiently restrictive to deliver an optimality condition in which an unbiased judge must equalize misconduct rates across marginal defendants of different races.

These findings raise the question of what additional restrictions are needed for the decision model to generate such an optimality condition and thereby restore logical validity to the proposed outcome test. In Section 4, we consider two such restrictions, and we show their strong implications for the admissible behavior of bail judges and defendants. Importantly, these restrictions and their implications are distinct from the issues with the outcome test that are discussed in ADY.

The first possible restriction is to exclude race from the expected cost function. This restriction restores the logical validity of the outcome test by assuming away any objective statistical relationship between race and potential misconduct, leaving subjective judge bias as the only explanation for any observed racial disparities in marginal defendant outcomes. To justify such a restriction, the analyst would need to not only rule out any direct causal effects of race on misconduct, but also require the judge's information set at the bail hearing to encompass all determinants of misconduct that correlate with race. These conditions seem unlikely to prevail in this setting and many others, given that a defendant's race (or other group membership of interest) is not plausibly randomly assigned relative to his own other characteristics, and the decision maker's information set only includes a limited subset of those characteristics.

Alternatively, on the other side of the cost-benefit model, the analyst might consider restricting the expected benefit of release to be invariant to all non-race defendant characteristics. This restriction restores the logical validity of the outcome test by eliminating all analyst-unobserved heterogeneity in how a given judge assesses the benefits of release across all the cases she adjudicates, leaving racial bias and the probability of misconduct as the sole sources of variation in a given judge's decisions. This restriction has strong implications for the admissible behavior of bail judges and defendants, as well as for the measurement of the misconduct outcome. To justify it, the analyst would need to assume that any racially biased judge is equally biased against all defendants of the same race, no matter their other analyst-unobserved characteristics like speech patterns, body weight, disabilities, socioeconomic status, and physical demeanor; that all possible biases in the

decision making of judges manifest themselves in only one form, as racial bias; that no errors in judge predictions of misconduct, nor in the analyst's measurement of misconduct, correlate with any non-race defendant characteristics; and that no judge considerations, other than misconduct probabilities, vary across defendants with different non-race characteristics.

It is important to observe that ADY do consider complications to the outcome test posed by measurement error, prediction error, and omitted judge considerations, but only to the extent that these confounding factors correlate with defendant race. One perhaps surprising lesson from our analysis is that these factors can invalidate the outcome test even if they are completely independent of race, or whichever characteristic the analyst is investigating as a potential target of bias.

In Section 5, we examine how the logical validity of the outcome test relates to its econometric viability in an instrumental variables framework. The empirical quantities of interest in the outcome test—in this case, misconduct rates of marginal black and marginal white defendants—correspond to marginal treatment effects in the framework of Heckman and Vytlacil (2005). In order to identify these MTEs, the analyst must impose the instrument monotonicity condition of Imbens and Angrist (1994), which requires uniformity across judges in how they treat different types of defendants. Without such cross-judge restrictions, it is not possible to infer characteristics of defendants at the margin of being released, including how likely these defendants are to commit pre-trial misconduct. We show, however, that these requirements for econometric viability of the outcome test are distinct from issues of logical validity. Importantly, imposing the restrictions required to identify MTEs does not restore the logical validity of the outcome test, meaning that the analyst may successfully recover race-specific marginal misconduct rates that are uninformative about judge racial bias. Conversely, we also show that restrictions that restore the logical validity of the test do not guarantee that its key components are empirically identified.

These results raise the questions: What kind of decision model enables both a logically valid and econometrically viable outcome test? What does this model imply about the permissible behavior of judges and defendants? In Section 6, we answer these questions. We begin by showing that a particular type of Roy selection model, known as the "extended" Roy model (e.g. Heckman and Vytlacil, 2007; D'Haultfœuille and Maurel, 2013), delivers both a logically valid and econometrically viable outcome test. This extended Roy model corresponds to ADY's general cost-benefit decision model subject to the exclusion of all non-race defendant characteristics from the expected benefit

function. On one hand, it is more flexible than the basic Roy (1951) model of perfect self-selection on potential outcomes because it allows for a decision threshold that varies across judges and defendant race. On the other hand, it is considerably more restrictive than the generalized Roy model, including the one ADY write down, and these additional restrictions are shown to have strong implications for the admissible behavior of judges and defendants.

Since the extended Roy model imposes additional restrictions beyond those required to identify outcomes of marginal defendants of each race, it generates clear testable implications. In Section 6, we explore these implications, showing that the race-specific MTE functions that identify marginal defendant misconduct rates must be monotonically increasing across their entire domain. Since these MTEs are the key empirical objects of interest in the outcome test, assessing whether they are monotonic—and thus assessing the validity of the extended Roy model—is a natural next step after their estimation. Interestingly, ADY estimate MTEs that do not appear to increase monotonically, raising concerns about whether bail judge behavior is consistent with the extended Roy model on which the logic of the outcome test rests.

We conclude in Section 7 with a discussion of the broader lessons that may be drawn from our study about the use of outcome tests for detecting bias in decision making. Most immediately, our analysis maps readily into the many decision environments that feature exogenously assigned decision makers, like bail judges, who vary progressively in their decision tendencies. A rapidly growing literature has employed these institutional features to identify causal consequences of an array of policy-relevant decisions, including criminal sentencing (Kling, 2006; Bhuller et al., 2020), welfare eligibility (Maestas et al., 2013; Dahl et al., 2014), patent protection (Galasso and Schankerman, 2015; Sampat and Williams, 2019), eminent domain (Belloni et al., 2012), foster care (Doyle, 2007, 2008), and bankruptcy protection (Dobbie and Song, 2015). An important innovation of ADY is their recognition that these institutional features also help address two key methodological challenges faced by the discrimination literature, notably selection and inframarginality.[2] Furthermore, decision makers in most of these environments are public officials who do not need to compete for cases, so there is little potential for market forces to erode the consequences of bias in equilibrium (e.g. Arrow, 1973).[3] And yet, as our analysis shows, key challenges remain to detecting bias with

[2]See Persico (2009) and Brock et al. (2012) for insightfully structured reviews of this literature, including influential contributions by Ayres and Waldfogel (1994), Knowles et al. (2001), and Anwar and Fang (2006).

[3]Lang and Lehmann (2012) review the theory and empirics of racial discrimination in labor markets.

outcome tests in these environments. The lessons we draw also apply beyond such environments, moreover, given that we identify challenges to detecting bias that arise in spite of, not because of, such methodologically amenable institutional features.

## 2 Framework

In this section, we lay out the framework for detecting bias in decision making. For concreteness, we consider the setting of ADY where bail judges decide whether to release or detain defendants who are awaiting trial. We discuss how our arguments generalize to a wide range of empirical settings in Section 7.

### Setup and Notation

Let $D$ indicate the decision of whether a bail judge releases a given defendant: $D = 1$ if the defendant is released, and $D = 0$ if the defendant is detained. The outcome of interest, $Y$, is a binary variable indicating whether the defendant commits any pre-trial misconduct ($Y = 1$) or not ($Y = 0$). Let $Y_1$ and $Y_0$ denote potential outcomes, where $Y_1$ is the potential outcome associated with release and $Y_0$ is the potential outcome associated with detention. The observed outcome is $Y = DY_1 + (1 - D)Y_0$.

Judges are indexed by $Z$. Defendants are characterized by their race $R \in \{w, b\}$, with $w$ denoting a white defendant and $b$ denoting a black defendant, as well as any non-race characteristics $V$ that are observable by the judge. The analyst observes $R$ but not $V$, whereas judges observe both $R$ and $V$. To simplify the notation, we abstract from defendant characteristics that are observable by both the judge and the analyst. Our analysis could easily proceed by conditioning on such characteristics and defining $V$ as the characteristics that remain unobservable to the analyst.

### Basic Assumptions

We denote by $P$ the distribution of $(Y_1, Y_0, Z, R, V)$. Throughout the paper, we make the following assumption on $P$:

**Assumption 2.1.** *The distribution $P$ satisfies*

*(i)* $(Y_1, Y_0, R, V) \perp\!\!\!\perp Z$.

*(ii) $Z$ is continuously distributed on $\mathcal{Z} \subseteq \mathbf{R}$ with density $f_Z(z)$.*

*(iii) $V$ is continuously distributed on the convex set $\mathcal{V} \subseteq \mathbf{R}$ with density $f_V(v)$.*

Assumption 2.1(i) reflects random assignment of judges to cases, along with an exclusion restriction that judges do not directly affect defendant potential outcomes. Assumption 2.1(ii) assumes a continuum of judges. Assumption 2.1(iii) assumes $V$ is a scalar random variable that takes values on a convex set and is continuously distributed.

These assumptions may be strong in some settings. Judges may specialize in certain types of cases, or there may be only a small number of judges, or non-race defendant characteristics may not be easily distilled into one dimension. We intentionally abstract from these issues to focus on problems that arise even in a best-case scenario for testing for bias in decision making. For the same reason, we assume that the analyst knows the population distribution of the observed data $(Y, D, R, Z)$ with certainty. This assumption clarifies that the problems we discuss should be viewed as both distinct from, and primary to, any issues of estimation and statistical inference.

### Decision Model

Following ADY, we assume that judge $Z$'s binary decision $D$ of whether to release a defendant of race $R$ with non-race characteristics $V$ can be modeled as the following cost-benefit comparison:

$$D = I\{\Lambda(R, V) \leq \tau(Z, R, V)\} \tag{1}$$

where $\Lambda(r, v)$ is referred to as the *expected cost* of release for a defendant of race $r$ and characteristics $v$, and $\tau(z, r, v)$ is the *expected benefit* of release by judge $z$ for a defendant of race $r$ and characteristics $v$.

As in ADY, we specify the expected cost of releasing a defendant as the expected causal effect of release on pre-trial misconduct, $\mathbb{E}[Y_1 - Y_0 | R, V]$. Because $Y$ is binary and defendants cannot commit misconduct if detained ($Y_0 \equiv 0$), the expected cost simplifies to the probability of misconduct if released:

**Assumption 2.2.** *The expected cost* $\Lambda$ *satisfies*

$$\Lambda(r, v) \equiv P\{Y_1 = 1 | R = r, V = v\} \ .$$

Note that our results hold regardless of whether $Y$ is binary or $Y_0 \equiv 0$; we could write Assumption 2.2 more generally as $\Lambda(R, V) \equiv \mathbb{E}[Y_1 - Y_0 | R, V]$. In some settings, decision makers may only care about one of these potential outcomes (e.g. $Y_1$), either because the other potential outcome (e.g. $Y_0$) does not vary across individuals, as in pre-trial release, or because the other potential outcome is not relevant for the decision maker's problem, as in the consumer lending setting of Dobbie et al. (2020). In other settings, like trial judges considering potential criminogenic effects of incarceration, both $Y_1$ and $Y_0$ may vary across individuals and be relevant for the decision maker, in which case the causal effect of treatment $\mathbb{E}[Y_1 - Y_0 | R, V]$ could be the quantity of interest. In any event, the main feature of Assumption 2.2 is that $\Lambda(\cdot)$ is tied to the analyst's outcome of interest $Y$.

The key content of the judge decision model in (1) can be summarized with four observations. The first is that the cost of release is tied to an outcome, defendant misconduct, that is observable by the analyst. By comparison, the benefit of release is unobserved by the analyst. Second, both race $R$ and non-race characteristics $V$ enter both the cost function and the benefit function. Third, the judge observes both $R$ and $V$, while the analyst observes $R$ but not $V$. Lastly, the expected cost of release $\Lambda(R, V)$ does not vary across judges $Z$. This reflects Assumption 2.1(i) that judges are randomly assigned to cases and do not directly affect defendant potential outcomes. By contrast, the perceived benefit of release $\tau(Z, R, V)$ is allowed to vary freely across judges $Z$.

### Definition of Racial Bias

Following ADY, we define racial bias as follows:

**Definition 2.1.** *We say judge* $z$ *is racially unbiased if* $\tau(z, r, v) = \tau(z, v)$ *for all* $v \in \mathcal{V}$. *If* $\tau(z, w, v) > \tau(z, b, v)$ *for all* $v \in \mathcal{V}$, *we say judge* $z$ *is racially biased against black defendants.*

This definition implies that a judge is racially biased against black defendants if she perceives greater benefit of releasing a white defendant than a black defendant if both have the same non-race characteristics $V$.

It is important to observe that this definition does not specify the motivations or mechanisms behind any bias in judge decision making. It could be that a judge has a personal taste for discrimination, or that her beliefs or optimization errors happen to favor particular groups at the expense of others, or that she uses race to statistically discriminate in terms of the expected benefit of release, which is unobserved by the analyst. All of these fall under the umbrella of racial bias according to Definition 2.1. We follow ADY and take this definition as given, recognizing that understanding the motives of discriminatory behavior can be important when considering potential policy responses.[4]

### Outcome Test of Racial Bias

To properly define the outcome test of racial bias, it is useful to make the following single-crossing assumption:

**Assumption 2.3.** *For all* $(z, r) \in \mathcal{Z} \times \{w, b\}$ *there exists* $V^*_{z,r} \in \text{int}(\mathcal{V})$ *such that*

$$\Lambda(r, V^*_{z,r}) = \tau(z, r, V^*_{z,r}) \ , \tag{2}$$

*and*

$$\begin{aligned} \Lambda(r, v) &< \tau(z, r, v) \quad \textit{for all } v < V^*_{z,r} \\ \Lambda(r, v) &> \tau(z, r, v) \quad \textit{for all } v > V^*_{z,r} \ . \end{aligned}$$

This assumption allows us to have a precise notion of *marginal* defendants, i.e. defendants for whom a given judge perceives exactly offsetting costs and benefits of release. In our notation, defendants of race $r$ with non-race characteristics equal to $V^*_{z,r}$ are marginal for judge $z$.

The outcome test proposed by ADY compares the pre-trial misconduct probability of judge $z$'s marginal white defendant with the pre-trial misconduct probability of her marginal black defendant. The test infers racial bias against black defendants if

$$\Lambda(w, V^*_{z,w}) > \Lambda(b, V^*_{z,b}) \ , \tag{3}$$

[4] Arnold et al. (2020) offer a recent exercise aiming to distinguish racial bias among New York City bail judges from statistical discrimination about misconduct rates.

and concludes there is no evidence of racial bias if

$$\Lambda(w, V^*_{z,w}) = \Lambda(b, V^*_{z,b}) \ . \tag{4}$$

In the next section we discuss whether the logic behind the outcome test is valid given the model of decision making in (1).

# 3 The Logical Invalidity of the Proposed Outcome Test

We now show that the outcome test of ADY, defined in (3) and (4), is logically invalid given their decision model in (1) and definition of racial bias in (5). This is done by generating broad classes of counterexamples which prove that the outcome test may conclude no bias even if each judge is racially biased, or conclude bias even if each judge is racially unbiased, or even conclude bias against one race when each judge is biased against the other race.

### Definition of a Logically Valid Outcome Test of Racial Bias

Before we present results on the logical invalidity of the proposed test, it is useful to make precise what it means for the outcome test to be logically valid:

**Definition 3.1.** *We say that the outcome test is logically valid if and only if*

$$\text{sign}\left(\Lambda(w, V^*_{z,w}) - \Lambda(b, V^*_{z,b})\right) = \text{sign}\left(\tau(z, w, v) - \tau(z, b, v)\right) \ \textit{for all } v \in \mathcal{V} \textit{ and } z \in \mathcal{Z} \ . \tag{5}$$

This definition implies that marginal white and marginal black defendants should have the same probability of pre-trial misconduct if and only if judge $z$ is racially unbiased. We view this as a minimal requirement for the outcome test in (3) and (4) to be a logically valid approach to detecting bias in the decision making of bail judges.

### Formal Results on Logical Invalidity

To generate clear and intuitive counterexamples of logical validity, it is useful to consider the subset of cost functions $\Lambda(\cdot)$ and benefit functions $\tau(\cdot)$ that satisfy the following conditions:

**Assumption 3.1.** *The expected cost function* $\Lambda(r, \cdot) : \mathcal{V} \to \mathcal{I} \subseteq \mathbf{R}$ *is continuous and weakly monotonically increasing in* $v$ *for all* $r \in \{w, b\}$.

**Assumption 3.2.** *The expected benefit function* $\tau(z, r, \cdot) : \mathcal{V} \to \mathcal{I} \subseteq \mathbf{R}$ *is continuous and strictly monotonically decreasing in* $v$ *for all* $z \in \mathcal{Z}$ *and* $r \in \{w, b\}$.

By focusing on functions that satisfy these simple conditions, it is straightforward to prove that the outcome test proposed in (3) and (4) is logically invalid per Definition 3.1, given the decision model in (1):

**Theorem 3.1.** *The following two results hold.*

*(i) Suppose that judge* $z \in \mathcal{Z}$ *is racially unbiased, i.e.*

$$\tau(z, r, v) = \tau(z, v) \text{ for all } v \in \mathcal{V} \ . \tag{6}$$

*Then, for any functions* $\Lambda(b, v)$ *and* $\tau(z, v)$ *jointly satisfying Assumptions 2.3-3.2 there exists a function* $\Lambda(w, v)$ *satisfying Assumptions 2.3-3.1 such that the marginal white defendant exhibits a higher misconduct probability than the marginal black defendant,*

$$\Lambda(w, V^*_{z,w}) > \Lambda(b, V^*_{z,b}) \ . \tag{7}$$

*(ii) Suppose that judge* $z \in \mathcal{Z}$ *discriminates against black defendants, i.e.*

$$\tau(z, w, v) > \tau(z, b, v) \text{ for all } v \in \mathcal{V} \ . \tag{8}$$

*Then, for any functions* $\Lambda(w, v)$ *and* $\tau(z, w, v)$ *jointly satisfying Assumptions 2.3-3.2 there exist functions* $\Lambda(b, v)$ *and* $\tau(z, b, v)$ *jointly satisfying Assumptions 2.3-3.2 such that the marginal white defendant exhibits a misconduct probability no larger than the marginal black defendant,*

$$\Lambda(w, V^*_{z,w}) \leq \Lambda(b, V^*_{z,b}) \ . \tag{9}$$

The theorem shows that the outcome test may conclude (i) bias even if each judge is racially unbiased, or (ii) no bias even if each judge is racially biased. The test may even conclude bias

against one race when in reality all judges are biased against the other race, as in part (ii) when (9) holds with strict inequality.

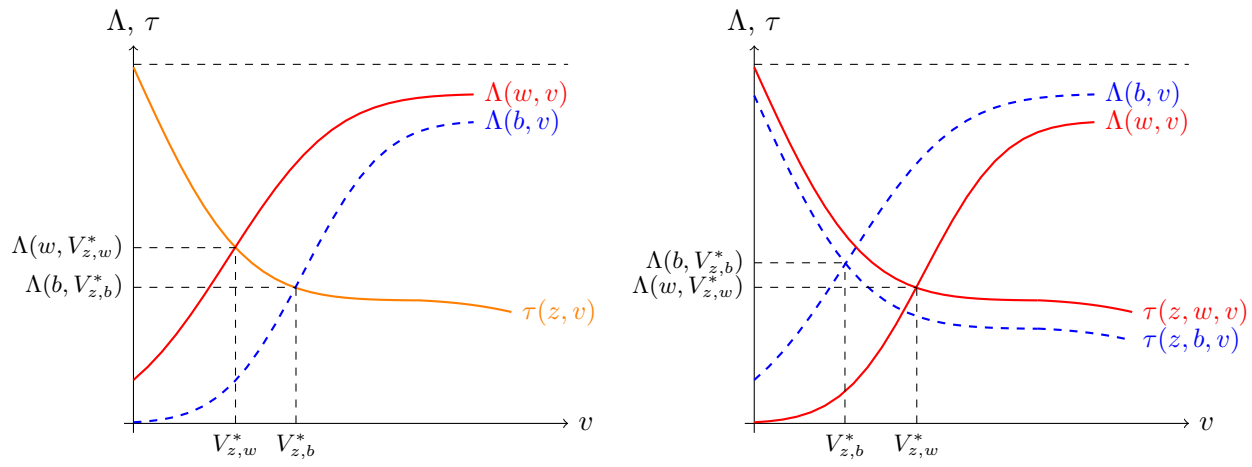
We relegate the proof of Theorem 3.1 to Appendix A and offer instead a graphical interpretation of its content in the next subsection. In this graphical analysis, it becomes evident why it is convenient to focus on the subset of cost and benefit functions that satisfy the monotonicity and continuity conditions of Assumptions 3.1 and 3.2. However, it is important to note that neither the monotonicity condition nor the continuity condition are required to prove that the outcome test is logically invalid given the model of decision making in (1). In Appendix A, we show that the conclusions of Theorem 3.1 continue to hold under weaker conditions that allow both functions to be non-monotonic and discontinuous.

### Graphical Representation of the Formal Results

In Figure 1, we explain the intuition behind the formal results of logical invalidity of the proposed outcome test by drawing examples of cost and benefit functions of black and white defendants facing a given judge. These functions jointly satisfy Assumptions 2.3-3.2. Thus, the functions we draw are illustrative cases of those that form the counterexamples in Theorem 3.1.

Figure 1a illustrates a case where the benefit function does not depend on race, $\tau(z, r, v) = \tau(z, v)$, so judge $z$ is racially unbiased according to Definition 2.1. However, the cost function $\Lambda(r, v)$ does vary with race, consistent with ADY's decision model in (1). In this case, $\Lambda(w, v)$ is greater than $\Lambda(b, v)$ for all $v$. So, even though judge $z$ is not biased against black defendants, the fact that black defendants have a lower probability of misconduct for a given value of the non-race characteristics $v$ leads to the marginal black defendant having a strictly lower misconduct probability than the marginal white defendant, $\Lambda(b, V^*_{z,b}) < \Lambda(w, V^*_{z,w})$. Thus, the cost and benefit functions that are drawn in Figure 1a illustrate a case in which both conditions (6) and (7) in result (i) of Theorem 3.1 hold. In other words, this counterexample shows that the outcome test may erroneously conclude bias when in fact judges are racially unbiased.

Figure 1b illustrates a case where race enters both the benefit function $\tau(z, r, v)$ and the cost function $\Lambda(r, v)$, again consistent with the decision model in (1). In this case, $\Lambda(b, v)$ is greater than $\Lambda(w, v)$ for all $v$, while $\tau(z, w, v)$ exceeds $\tau(z, b, v)$ for every $v$. So, even though judge $z$ discriminates against black defendants, the fact that black defendants have sufficiently higher probability of




(a) Intuition behind part (i)

(b) Intuition behind part (ii)

Figure 1: Intuition behind Theorem 3.1

misconduct for a given value of the non-race characteristics $v$ leads to the marginal black defendant having a strictly higher misconduct probability than the marginal white defendant, $\Lambda(b, V^*_{z,b}) > \Lambda(w, V^*_{z,w})$. Thus, the cost and benefit functions that are drawn in Figure 1b illustrate a case in which both conditions (8) and (9) in result (ii) of Theorem 3.1 hold. In other words, this counterexample shows that the outcome test may erroneously conclude bias against one race when in fact judges are racially biased against the other race.

The logical invalidity of the outcome test arises because the decision model of ADY in (1) does not deliver an optimality condition where an unbiased judge must equalize the probability of misconduct across marginal white and marginal black defendants. This raises the question of what additional restrictions are needed for the model to generate such an optimality condition and restore the validity of the outcome test.

The examples in Figures 1a and 1b point to two possible restrictions. On the one hand, Figure 1a suggests the possibility of excluding race from the cost function. Under this additional restriction of the model, the marginal white defendant would have the same misconduct probability as the marginal black defendant, and the outcome test would correctly conclude no bias. On the other hand, Figure 1b suggests the possibility of excluding non-race defendant characteristics $V$ from the benefit function. Under this additional restriction of the model, the marginal black defendant would have a lower misconduct probability than the marginal white defendant, and the outcome test would correctly conclude bias against black defendants. Motivated by the examples in Figures

1a and 1b, we next explore these restrictions in more detail and discuss their implications.

# 4 Restrictions That Restore the Logical Validity of the Test

We now consider two restrictions to the decision model in (1) that may help nullify the results in Theorem 3.1 and restore logical validity to the outcome test. One of these is to exclude race from the cost function. The other is to exclude non-race defendant characteristics from the benefit function. We now show the consequences of these restrictions both for the validity of the outcome test and for the admissible behavior of bail judges and defendants.

## 4.1 Excluding Race from the Cost Function

We first consider result (i) of Theorem 3.1. To overturn this result, it is enough to exclude race from the cost function:

$$\Lambda(r, v) = \Lambda(v) \text{ for all } v \in \mathcal{V} \ . \tag{10}$$

Under this restriction, the outcome test would conclude no bias if judges indeed are unbiased. Figure 1a illustrates this point. If both the cost function and the benefit function are independent of race, then their single intersection point must be the same for both races, and as such the marginal white defendant must have the same misconduct probability as the marginal black defendant.

To overturn result (ii) of Theorem 3.1, it is not sufficient to exclude race from the cost function. An additional restriction is needed. This is evident from Figure A.3a, which illustrates a case where race is excluded from the cost function, $\Lambda(r, v) = \Lambda(v)$, but enters the benefit function $\tau(z, r, v)$ such that judge $z$ is biased against black defendants. Since Assumption 3.1 allows the cost function to be only weakly monotonic, it is still possible for the misconduct probability to be the same for the marginal black defendant as for the marginal white defendant. Thus, even when judges are biased against black defendants, the outcome test could suggest no bias under the restriction in (10). This finding motivates Lemma 4.1, which shows that the logical validity of the outcome test is fully restored if the cost function is both independent of race and strictly monotonic:

**Lemma 4.1.** *Let Assumption 2.3 hold. Assume that the expected cost function satisfies* (10) *and is continuous and strictly monotonically increasing. Then, the outcome test is logically valid in the*

*sense of Definition 3.1.*

**Implications of This Restriction**

The restriction in (10) implies that conditioning on non-race defendant characteristics $V$ removes any statistical relationship between race $R$ and the probability of misconduct $\Lambda(\cdot)$. In addition to ruling out any direct causal effects of race on misconduct, this restriction requires the judge's information set at the bail hearing to encompass all determinants of misconduct that correlate with race, such that, conditional on this information set, race is as good as randomly assigned relative to all remaining unobserved factors that drive misconduct.

These conditions are unlikely to prevail in practice. As emphasized by Dobbie et al. (2018), "judges generally have limited information on which to base their decisions." Bail hearings tend to occur within a day or two of arrest and last just a few minutes. The bail judge's information set is typically limited to the current allegations against the defendant, any preliminary evidence available, the defendant's criminal record, perhaps some coarse measures of the defendant's employment status and living situation, and the defendant's observable demeanor during the brief hearing, which is conducted via videoconference in most jurisdictions (Arnold et al., 2018). The bail hearing is therefore unlikely to reveal all determinants of pre-trial misconduct that correlate with race, leaving ample scope for omitted variables like resources, motives, and networks that are unlikely to be statistically independent of race. Intuitively, while judges are plausibly randomly assigned to defendants, a defendant's race is not plausibly randomly assigned relative to all of his other characteristics unobserved by the judge. As a result, it seems difficult to justify excluding race from the expected cost function to satisfy the conditions of Lemma 4.1.

## 4.2 Excluding Non-Race Characteristics from the Benefit Function

Given the difficulty of justifying restriction (10) on the cost function, it may be tempting to instead make the following restriction on the benefit function:

$$\tau(z, r, v) = \tau(z, r) \text{ for all } r \in \{w, b\} \ . \tag{11}$$

Under this restriction, which removes any dependence of the benefit function on non-race defendant characteristics $V$, neither part (i) or part (ii) of Theorem 3.1 would hold. The validity of the outcome test is then fully restored as it would correctly conclude (no) bias if judges are (un)biased. The intuition for this follows from Figure 1a and 1b, since both of those results require $\tau(z, r, v)$ to vary by non-race characteristics. If not, the expected benefit of release for all defendants (including the marginal) of a given race facing a given judge is determined entirely by a fixed quantity. If this fixed benefit is higher (the same) for whites as compared to blacks, then its intersection with the white cost function, and thus the marginal white defendant's misconduct probability, must be higher (the same) as compared to the marginal black defendant. This argument is formalized in Lemma 4.2, which shows that the logical validity of the outcome test is fully restored if the benefit function is independent of non-race characteristics:

**Lemma 4.2.** *Let Assumption 2.3 hold. Assume that the expected benefit function satisfies* (11)*. Then, the outcome test is logically valid in the sense of Definition 3.1.*

**Implications of This Restriction**

Restriction (11) implies that each judge acts as if all defendants of the same race have exactly the same benefit of release. In other words, perceived benefits of release may vary across judges, but for a given judge, the benefit of release must be identical for all defendants of the same race, no matter their non-race characteristics. We now discuss five key implications of this restriction for the admissible behavior of bail judges and defendants, as well as for the measurement of the misconduct outcome.

First, restriction (11) implies that all bias in the decision making of bail judges must manifest itself in only one form, as racial bias. All other forms of bias are ruled out by (11), since the only admissible variation in a given judge's benefit function is by race, which constitutes racial bias per Definition 2.1. In reality, judges may discriminate against defendants across a wide range of non-race characteristics typically observable in a bail hearing, including physical appearances (Hatfield and Sprecher, 1986; Schneider, 2005), speech patterns (Bouchard Ryan et al., 1977; Gluszek and Dovidio, 2010), body weight (Brownell et al., 2005), age (Nelson, 2004), poverty status (Cozzarelli et al., 2001), disability (Dunn, 2014), and country of origin (Wagner et al., 2010), to name a few

(Dovidio et al., 2010). ADY discuss the possibility that judges have other forms of biases, arguing that in such cases their measure of "racial bias includes not only any bias due to phenotype, but also bias due to seemingly non-race factors that are correlated with, if not driven by, race." While we do not necessarily object to this broadening of the definition of racial bias, it is important to observe that such broadening does not solve the problem at hand. Crucially, the presence of judge biases with respect to any non-race defendant characteristics can invalidate the outcome test for racial bias even if those characteristics do not correlate with race or interact with race in judge decision making. The reason is that such biases must manifest themselves in the decision model (1) as a benefit function $\tau(\cdot)$ that varies with $V$. If $V$ represents body weight, for example, then a judge biased against overweight defendants must have $\tau(z, r, v) < \tau(z, r, v')$ for $v > v'$. Regardless of whether $V$ enters $\tau(\cdot)$ as additively separable and independent from $R$—i.e. even if the benefit function varies with non-race characteristics in exactly the same way for black and white defendants—it will violate (11) and therefore re-open the applicability of Theorem 3.1.

A related implication of restriction (11) is that any racially biased judge must be equally biased against all defendants of the same race. This homogeneity restriction is at odds with a large body of empirical evidence suggesting that prejudices specifically against blacks can vary considerably in magnitude with a wide range of characteristics, including skin tone (Maddox, 2004; Hagiwara et al., 2012), facial features (Livingston and Brewer, 2002), speech patterns (Tucker and Lambert, 1969; Koch et al., 2001), age (Kang and Chasteen, 2009), height (Hester and Gray, 2018), weight (Hebl and Turchin, 2005), gender (Ghavami and Peplau, 2012), and other dimensions (Kang and Bodenhausen, 2015). Many of these characteristics may be unobserved by the analyst, yet fully or partially observed by the judge, and therefore represented in $V$. This makes the a priori exclusion of $V$ from the bias-governing benefit function in (11) difficult to justify, especially in a model of decision making specifically employed to measure racial bias.

A third implication of restriction (11) is that any measurement error in the outcome variable (pre-trial misconduct) must be uncorrelated with non-race defendant characteristics $V$. To see this, note that the "benefit" function $\tau(\cdot)$ is ultimately a catch-all for all residual factors in the decision model in (1) beyond the probability of misconduct *as measured in the data*. ADY discuss the possibility of measurement error, noting that "our test for racial bias assumes that any measurement error in the outcome is uncorrelated with race. This assumption would be violated if, for example,

judges minimize new crime, not just new arrests, and the police are more likely to rearrest black defendants conditional on having committed a new crime... In this scenario, we overestimate the probability of pre-trial misconduct for black versus white defendants at the margin and, as a result, underestimate the true amount of racial bias in bail setting." While this may be true, the implications of restriction (11) are actually much stronger: this restriction can fail even if measurement error in misconduct is independent of race $R$. What also matters for the validity of the outcome test is whether measurement error is correlated with any non-race characteristics $V$. Defendants with more criminal experience, for example, regardless of their race, may be more likely to commit pre-trial misconduct that goes undetected and thus unmeasured. Such measurement error would necessarily be folded into $\tau(\cdot)$ given the decision model in (1), inducing dependence of $\tau(\cdot)$ on $V$ and thus violating (11).

A fourth implication of restriction (11) is that judges' accuracy in predicting pre-trial misconduct cannot vary systematically with any non-race defendant characteristics $V$. This implication is at odds with recent work comparing release decisions of bail judges with machine learning predictions of defendant misconduct. For example, Kleinberg et al. (2018) find that judges appear to systematically underestimate the misconduct of defendants who happen to face a minor current charge but have other characteristics (e.g. prior convictions) that strongly predict misconduct. Conversely, judges appear to systematically overestimate the misconduct of low-risk defendants who happen to face a more serious current charge. The results in Kleinberg et al. (2018) also suggest that judges may place excessive weight on salient interpersonal factors unobserved by the analyst, like a defendant's attitude and physical appearance during the hearing, when predicting misconduct. ADY recognize that prediction errors are likely in this setting, given that "bail judges must make quick judgments on the basis of limited information, with virtually no training and, in many jurisdictions, little experience working in the bail system." However, ADY argue that racially biased prediction errors can simply be viewed as one of many sources of racial bias. While such an interpretation may be natural, it does not relax the strong implication of restriction (11) for the accuracy of judge predictions. This is because restriction (11) can fail even if prediction errors are independent of race $R$. What matters for the validity of the outcome test is whether such errors vary with non-race characteristics $V$, which seems difficult to rule out, especially given the results of Kleinberg et al. (2018).

Lastly, restriction (11) implies that the probability of pre-trial misconduct is the only judge consideration in the release decision that may vary in magnitude across defendants with different non-race characteristics. In reality, bail judges in most jurisdictions are instructed to gauge not only the likelihood of any misconduct occurring but also its potential severity along several dimensions, including risks to the public at large, danger to specific victims and witnesses, and challenges to the integrity of the judicial process (American Bar Association, 2007). Furthermore, some judges may weigh non-misconduct consequences of releasing rather than detaining a defendant, including job and family stability, as well as decreased likelihood of conviction and future recidivism (e.g. Leslie and Pope, 2017; Dobbie et al., 2018), which Kleinberg et al. (2018) refer to as "omitted payoffs." ADY argue that their estimates will suffer from "omitted payoff bias if, for example, bail judges consider how pre-trial detention impacts a defendant's employment status and this outcome is correlated with race." In fact, restriction (11) shows that the logical validity of the outcome test also hinges on whether any of these judge considerations vary with non-race defendant characteristics $V$. Notably, in predicting misconduct severity (conditional on its probability), bail judges are explicitly instructed to use non-race predictors like the defendant's criminal record, employment status, and family structure.[5] Non-misconduct "omitted payoffs" of release are also likely to vary substantially with defendant characteristics: family disruption, for example, is less relevant for a single young defendant than a married one with several children, and Dobbie et al. (2018) find evidence that the employment consequences of pre-trial detention are larger for defendants with less criminal history. Regardless of whether bail judges are officially instructed to incorporate or ignore any of these considerations beyond the probability of misconduct, if any judges do so in practice, then the outcome test may cease to be informative about the magnitude, or even the sign, of any racial bias that judges may also act upon.

To critically assess the plausibility of restriction (11), it is important to observe that, like Tolstoy's happy and unhappy families, this restriction can fail for many reasons but hold in only one way, by avoiding each of the reasons for failure. In other words, to restore the logical validity of the outcome test by invoking restriction (11), it is necessary to argue no forms of bias other than racial, no heterogeneity in racial bias across other defendant characteristics, no errors in data collection or judge predictions that correlate with non-race defendant characteristics, and no judge

[5]See, for example, 18 U.S. Code § 3142(g) and California Penal Code 1275(a)(1).

considerations beyond the probability of misconduct that vary across defendants with different non-race characteristics.

# 5 Econometric Viability of the Outcome Test

The outcome test of racial bias defined in (3) and (4) requires the analyst to compare the misconduct probabilities of marginal white and marginal black defendants. To do so, however, it is necessary to first recover these quantities from the distribution of the observed data $(Y, D, R, Z)$. Without restrictions other than those embedded in the decision model (1), it is not possible to recover misconduct probabilities of marginal defendants. This is because the model restricts the behavior of a given judge, but it does not impose restrictions on how judges differ in their behavior. Without cross-judge restrictions, it is not possible to isolate characteristics of defendants at the margin of being released, including how likely these defendants are to commit pre-trial misconduct.

This observation motivates the analysis in this section, where we consider additional assumptions that would identify the misconduct probabilities of marginal defendants through the marginal treatment effects (MTE) framework developed by Heckman and Vytlacil (2005) and applied by ADY. We then show that such restrictions do not restore the logical validity of the outcome test, and conversely, restrictions that restore the logical validity of the test do not necessarily lead to identification of misconduct probabilities of marginal defendants via MTEs.

## 5.1 Marginal Misconduct Probabilities as MTEs

We now study identification of the misconduct probabilities of marginal defendants in the marginal treatment effects (MTE) framework of Heckman and Vytlacil (2005). We focus on identifying MTEs with a continuous instrument as a best-case scenario, since discrete instruments and the identification of local average treatment effects (LATEs) would only introduce additional inframarginality complications when attempting to isolate marginal defendants, as ADY acknowledge.[6]

In order for the decision model in (1) to admit a valid MTE representation, one needs to argue

[6]See Brinch et al. (2017) and Mogstad et al. (2018) for methods to recover MTEs with discrete instruments.

that the model in (1) can be equivalently represented as

$$D = I\{U_r \leq p(z,r)\} \ , \tag{12}$$

where $U_r$ is a uniformly distributed random variable on $[0,1]$, unobserved by the analyst, and $p(z,r)$ is the propensity score (probability of release) for a defendant of race $r$ facing judge $z$. This representation allows the analyst to define the race-specific MTE functions as

$$MTE_r(u) \equiv \mathbb{E}[Y_1 - Y_0 | U_r = u, R = r] = P\{Y_1 | U_r = u, R = r\} \ , \tag{13}$$

where the second equation incorporates the fact that $Y$ is binary and $Y_0 = 0$ for all defendants. Since marginal defendants are those with $U_r = p(z,r)$ according to (12), it follows immediately from (13) that the misconduct rate of marginal defendants of race $r$ facing judge $z$ is identified by $MTE_r(p(z,r))$—that is, the marginal treatment effect function evaluated at the propensity score.

A commonly used approach to derive the representation in (12) from an economic decision model is to apply the probability integral transform, which consists of applying a properly chosen CDF to both sides of the arguments entering the indicator function. Despite being a popular and simple manipulation, this approach requires the original decision model to satisfy a certain structure that the judge decision model in (1), unfortunately, does not satisfy without further assumptions. To see this, let

$$F_{\Lambda|R=r}(x) \equiv P\{\Lambda(R,V) \leq x | R = r\}$$

denote the CDF of $\Lambda$ conditional on $R = r$. Applying this CDF to both sides of (1) yields

$$\begin{aligned} D &= I\{\Lambda(r,V) \leq \tau(z,r,V)\} \\ &= I\{F_{\Lambda|R=r}(\Lambda(r,V)) \leq F_{\Lambda|R=r}(\tau(z,r,V))\} \\ &= I\{U_r \leq F_{\Lambda|R=r}(\tau(z,r,V))\} \ , \end{aligned}$$

where the marginal distribution of $U_r$ is uniform on $[0,1]$. However, in this case

$$P\{D = 1 | Z = z, R = r\} = P\{U_r \leq F_{\Lambda|R=r}(\tau(z,r,V))\} \neq F_{\Lambda|R=r}(\tau(z,r,V)) \ , \tag{14}$$

since the term $F_{\Lambda|R=r}(\tau(z,r,V))$ is random and correlated with $U$ by virtue of being a non-trivial function of $V$, which is unobserved by the econometrician and not separable from the instrument $Z$. It follows that $F_{\Lambda|R=r}(\tau(z,r,V))$ is not a propensity score, and the treatment model in (1) cannot be taken to the MTE framework using the probability integral transform without further restrictions.

**Identification under Monotonicity**

The identification argument of ADY relies on the assumption that the instrument $Z$ satisfies the independence, exclusion, and monotonicity assumptions of Imbens and Angrist (1994). Our Assumption 2.1 already imposed instrument independence and exclusion: judges are randomly assigned to defendants, and judges have no direct effects on misconduct beyond the pre-trial release decision. Assumption 2.1 did not, however, impose the monotonicity condition of Imbens and Angrist (1994), since this condition restricts behavior across judges and our results up to this point have only concerned the behavior of a given judge.

Specifically, monotonicity in the context of the decision model in (1) requires that exogenous reassignment from one judge to another weakly increases the benefit function (and thus the likelihood of release) for every defendant of a given race, or weakly decreases it for every defendant of a given race:

$$\tau(z,r,v) \geq \tau(z',r,v) \text{ for all } v \in \mathcal{V} \quad \text{or} \quad \tau(z,r,v) \leq \tau(z',r,v) \text{ for all } v \in \mathcal{V}\,, \tag{15}$$

for any two judges $z$ and $z'$ within each $r \in \{w,b\}$. Adding this monotonicity restriction to the decision model in (1) is enough to obtain a valid MTE representation, since the equivalence result of Vytlacil (2002) guarantees the existence of a weakly separable selection model of treatment assignment as in (12). We note, however, that this result does not follow from applying the probability integral transform, as the problems described in (14) can still hold for benefit functions that satisfy (15). This contrasts with ADY's Appendix Equation (54), where a probability integral transform is erroneously used to justify the existence of an MTE representation of (1).

### 5.2 An MTE Representation Does Not Imply a Logically Valid Outcome Test

We now show that even after imposing restrictions that enable the econometric viability of the outcome test, this does not necessarily solve the issues of logical validity discussed earlier. The reason is that logical validity depends on the set of restrictions imposed on the behavior of a given judge, while econometric viability depends on restrictions imposed on behavior across judges.

A simple restriction that would guarantee instrument monotonicity and an MTE representation of the judge decision model in (1) is additive or multiplicative separability between $z$ and $v$ in the benefit function $\tau(\cdot)$. Consider, for example, the case where the benefit function is additively separable in $z$ and $v$,

$$\tau(z, r, v) = \tau_1(z, r) + \tau_2(v, r) \ , \tag{16}$$

for some functions $\tau_1(\cdot)$ and $\tau_2(\cdot)$. This restriction not only implies that the monotonicity restriction in (15) holds, but it also allows for a derivation of (12) via the probability integral transform. To see this, let $\Lambda^*(R, V) \equiv P\{Y_1 = 1|R, V\} - \tau_2(V, R)$ and let $F_{\Lambda^*|R=r}(\cdot)$ denote the CDF of $\Lambda^*$ conditional on $R = r$. We can now rewrite the decision model in (1) as

$$\begin{aligned} D &= I\{\Lambda^*(r, V) \leq \tau_1(z, r)\} \\ &= I\{F_{\Lambda^*|R=r}(\Lambda^*(r, V)) \leq F_{\Lambda^*|R=r}(\tau_1(z, r))\} \\ &= I\{U_r \leq p(z, r)\} \ . \end{aligned}$$

Since $U_r$ is a uniformly distributed random variable on $[0, 1]$ and $p(z, r) \equiv F_{\Lambda^*|R=r}(\tau_1(z, r))$ is the propensity score, this model is isomorphic to (12) and thus has a valid MTE representation.

While separability of $\tau(z, r, v)$ in $z$ and $v$ is therefore sufficient to enable econometric viability, this restriction does not restore the logical validity of the outcome test. To see this, simply observe that the separability restriction in (16) does not remove the dependence of the benefit function on $V$, allowing it to still satisfy the assumptions behind the counterexamples in Theorem 3.1. Thus, the outcome test proposed by ADY remains logically invalid even after imposing additional restrictions to the judge decision model that enable identification of marginal misconduct rates.

### 5.3 A Logically Valid Outcome Test Does Not Imply an MTE Representation

In the other direction, we conclude this section by showing that restrictions that restore the logical validity of the outcome test do not necessarily make it econometrically viable. To see this, consider the restriction in Section 4.1, where the cost function $\Lambda(\cdot)$ is independent of $R$ and strictly monotonic in $V$. Lemma 4.1 showed that this restriction restores the logical validity of the outcome test. This restriction does not imply an MTE representation, however, since it leaves the benefit function $\tau(z, r, v)$ unrestricted across judges and free to violate the monotonicity assumption in (15).

## 6 The Extended Roy Model Delivers a Valid Outcome Test

In the previous section, we showed that restrictions that enable econometric viability of the outcome test do not necessarily restore its logical validity, and conversely, restrictions that restore logical validity to the outcome test do not necessarily make it econometrically viable. These results raise the questions: What kind of decision model enables both a logically valid and econometrically viable outcome test? What does this model imply about the permissible behavior of judges and defendants and the data generating process? We now answer these questions, and by doing so, offer a blueprint for researchers considering the use of outcome tests to detect bias across a wide range of empirical environments, as discussed in greater detail in Section 7.

### 6.1 Restriction that Restores Both Logical Validity and Econometric Viability

Consider again the restriction in Section 4.2 that the expected benefit of releasing a defendant does not depend on his non-race characteristics $V$. Under this restriction, the decision model in (1) simplifies to

$$D = I\{\Lambda(R, V) \leq \tau(Z, R)\} \ . \tag{17}$$

In this restricted decision model, each judge sets a fixed threshold of tolerable misconduct for all defendants of the same race: every defendant with a probability of misconduct below this threshold is released, and all those above it are detained. This threshold may differ for white versus black defendants, i.e. $\tau(z, w) \neq \tau(z, b)$, which is how racial bias of judge $z$, as defined in Definition 2.1, would manifest itself in this restricted model.

As shown in Lemma 4.2, the exclusion of non-race characteristics from the benefit function in (17) yields a logically valid outcome test. Since marginal defendants of each race are defined by $\Lambda(r, V^*_{z,r}) = \tau(z, r)$ in this restricted model, we necessarily have

$$\Lambda(w, V^*_{z,w}) - \Lambda(b, V^*_{z,b}) = \tau(z, w) - \tau(z, b) \ ,$$

which satisfies Definition 3.1 and implies the outcome test will correctly conclude (no) bias if judge $z$ is (un)biased.

Furthermore, the restricted decision model in (17) has a valid MTE representation, which means that marginal misconduct rates of white and black defendants can be identified as race-specific MTEs. This follows immediately after noticing that the restriction on the benefit function in (17) is equivalent to the separability condition in (16) with $\tau_2(v, r) \equiv 0$, and so the monotonicity condition in (15) necessarily holds. The propensity score in this model is given by

$$p(z, r) = F_{\Lambda|R=r}(\tau(z, r)) \ , \tag{18}$$

where $F_{\Lambda|R=r}(\cdot)$ is again the CDF of $\Lambda$ conditional on $R = r$.

### 6.2 The Restriction Delivers an Extended Roy Model

The decision model in (17) is a type of Roy model, with potential outcomes playing a primary role in the treatment decision. As in Heckman and Vytlacil (2007) and D'Haultfœuille and Maurel (2013), we will refer to this model as the "extended" Roy model. It is a compromise between two extremes on the Roy spectrum: the basic Roy (1951) model of perfect self-selection, and the flexible generalized Roy model of Björklund and Moffitt (1987) and Heckman and Vytlacil (2005).

In the basic Roy model, decision makers are assumed to directly maximize observable outcomes, implying that they choose treatment if and only if the outcome with treatment exceeds the outcome without treatment. The model in (17) extends the basic Roy model by allowing for a decision threshold that varies with non-outcome variables observable to the analyst, namely the judge assignment $z$ and defendant race $r$. Within each judge and defendant race group, however—i.e. conditional on analyst observables—it is important to note that the sole source of variation that

determines the treatment decision $D$ is variation in the expectation of the defendant potential outcome $Y_1$.

By comparison, the generalized Roy model significantly broadens the Roy framework by allowing treatment decisions to depend on non-outcome considerations that are heterogeneous across individuals and unobservable to the analyst. The generalized Roy model thus allows, but does not impose, a relationship between treatment assignment and potential outcomes, and is therefore flexible enough to encompass most modern empirical work using instrumental variables to identify treatment effects (Heckman, 2010). In fact, ADY's judge decision model in (1) can be characterized as a generalized Roy model, since a benefit function $\tau(\cdot)$ that depends on $V$ comprises a component of the decision model that is decoupled from the observable outcome and varies in ways that are unobservable to the analyst. Our results on logical invalidity therefore demonstrate that outcome tests based on generalized Roy decision models like (1) will not be logically valid without further restrictions.

While the extended Roy model in (17) is somewhat more flexible than the basic Roy model, it is important to observe that it is considerably more restrictive than the generalized Roy model, and that these additional restrictions have strong implications for the admissible behavior of judges and defendants, as well as the analyst's measurement of the outcome. As discussed in Section 4.2, to justify this extended Roy model's exclusion of non-race characteristics $V$ from the benefit function, it is necessary to argue no forms of judge bias other than racial, no variation in the strength of a judge's racial bias across other defendant characteristics, no errors in judge predictions or the measurement of the outcome which correlate with non-race defendant characteristics, and no other judge considerations in the release decision that vary with defendants' non-race characteristics aside from the probability of any pre-trial misconduct occurring.

### 6.3 Testable Implications of the Extended Roy Model

Because the extended Roy model in (17) imposes additional restrictions beyond those required to identify treatment effects, it generates testable implications beyond those of the LATE/MTE framework.[7] In particular, the race-specific MTE functions that identify marginal defendant misconduct rates must be monotonically increasing across the entire support of the propensity score.

[7] See Mogstad et al. (2018) and the references therein for testable implications of the LATE/MTE framework.

Intuitively, a judge with a relatively high propensity score in the extended Roy model must be a judge with a relatively high fixed threshold of tolerable misconduct, since variation in potential misconduct is the only variation that determines whether defendants of a given race are released by a given judge. Since marginal defendants are those with misconduct probabilities equal to this tolerance threshold, the functions $MTE_r(p(z,r))$ that identify marginal misconduct rates of defendants of race $r$ facing judge $z$ must be increasing in the judge propensity score $p$.

To show this formally, consider again the race-specific MTE function defined in (13). If the CDF of $\Lambda(r,V)$ conditional on $R=r$, $F_{\Lambda|R=r}(\cdot)$, is one-to-one for each $r$, then evaluating $MTE_r(u)$ at $u=p(z,r)$ yields

$$\begin{aligned} MTE_r(p(z,r)) &= \mathbb{E}[Y_1|U_r = p(z,r), R=r] \\ &= \mathbb{E}[Y_1|F_{\Lambda|R=r}(\Lambda(r,V)) = F_{\Lambda|R=r}(\tau(z,r)), R=r] \\ &= \mathbb{E}[Y_1|E[Y_1|r,V] = \tau(z,r), R=r] \\ &= \tau(z,r) \\ &= F^{-1}_{\Lambda|R=r}(p(z,r)) \ , \end{aligned} \tag{19}$$

where the second equality follows from the equivalence of the extended Roy model in (17) and $D = I\{U_r \leq p(z,r)\}$, the third equality follows from Assumption 2.2, the fourth equality follows from the properties of conditional expectations, and the last equality follows from (18).

It follows from (19) that $MTE_r(p(z,r))$ must be monotonically increasing in the propensity score $p$ for each $r \in \{w,b\}$. This is a testable implication of the extended Roy model in (17). Since $MTE_w(\cdot)$ and $MTE_b(\cdot)$ are the key empirical objects of interest in the outcome test, assessing whether they are monotonically increasing—and thus assessing the validity of the extended Roy model—is a natural next step after their estimation. ADY do not formally test for increasing MTEs in their empirical results, but their Figure II shows a roughly flat $MTE_b(\cdot)$ for black defendants and perhaps even a (noisily) decreasing $MTE_w(\cdot)$ for white defendants. Such results are suggestive of an empirical rejection of the extended Roy model on which the logic of the outcome test rests.

# 7 Broader Lessons for Detecting Bias in Decision Making

For concreteness, we have conducted our analysis within the pre-trial release environment of ADY, as it offers an illuminating case study of the challenges that arise, even after important innovations, in inferring bias from outcome tests. But the lessons we draw extend to settings well beyond bail.

Most immediately, our analysis readily applies to the wide array of decision environments that share institutional features in common with pre-trial release. In our notation, these features include exogenously assigned decision makers $Z$, who vary progressively in their decision $D$ tendencies, and face individuals characterized by their group membership $R$ potentially subject to bias (e.g. race, gender, national origin), analyst-unobserved other characteristics $V$, and analyst-observed outcomes $Y$. Researchers are increasingly interested in these environments as laboratories for causal inference about the consequences of policy-relevant decisions, including criminal sentencing (Kling, 2006; Bhuller et al., 2020), welfare eligibility (Maestas et al., 2013; Dahl et al., 2014), patent protection (Galasso and Schankerman, 2015; Sampat and Williams, 2019), eminent domain (Belloni et al., 2012), foster care (Doyle, 2007, 2008), and bankruptcy protection (Dobbie and Song, 2015). As ADY observe, these environments are also potentially attractive testing grounds for bias in decision making, since exogenous assignment and progressively varying decision tendencies may help address some perennial methodological challenges in the literature on discrimination, namely selection and inframarginality.

Our analysis also applies more broadly, however, given that we highlight challenges to detecting bias that arise in spite of, not because of, such methodologically amenable institutional features. Addressing selection and inframarginality is not sufficient for a valid outcome test. Regardless of whether a given decision environment features exogenous assignment and progressive variation in decision tendencies, or whether a researcher develops other methods of dealing with selection and inframarginality, the following fundamental point remains. Proposing a valid outcome test for bias requires specifying a model of decision making that is restrictive enough to deliver empirically distinguishable characterizations of biased versus unbiased behavior, but rich enough to incorporate the essential elements of the optimization problem faced by the relevant decision makers.

As we show in this paper, this trade-off manifests itself in the choice between alternative versions

of the seminal Roy model of selection.[8] At one extreme, the basic Roy (1951) model assumes that decision makers simply maximize the outcome observed by the analyst. In the classic setting of occupational choice, for example, a worker selects the occupation in which she earns the highest wage, with no other considerations. At the other extreme, the generalized Roy model (Björklund and Moffitt, 1987; Heckman and Vytlacil, 2005) makes few restrictions on why and how choices are made. In it, a decision maker chooses treatment if the subjective surplus from doing so is positive, such that the perceived benefit outweighs the perceived cost. In occupational choice, this model allows a worker to consider not only the potential wage of an occupation, but also non-pecuniary costs and benefits of taking it up, which may relate to unobserved preferences, constraints, and abilities that vary across workers. In ADY's model of pre-trial release, a bail judge considers not only potential misconduct, but also benefits of release that vary across defendant characteristics unobserved by the analyst. Due to this flexibility of the generalized Roy model, it does not yield a valid outcome test of bias, since it does not require an unbiased decision maker to equalize the outcomes of marginal individuals across different groups.

A middle ground is offered by the so-called extended Roy model (e.g. Heckman and Vytlacil, 2007; D'Haultfœuille and Maurel, 2013). It delivers both a logically valid and econometrically viable outcome test of decision maker bias. As in the generalized Roy model, the extended Roy model assumes that a decision maker chooses treatment if the subjective surplus from doing so is positive, and this surplus can depend on other considerations besides the analyst's outcome of interest. However, the extended Roy model confines all such non-outcome considerations into an index that only varies with analyst observables, like the identity of the decision maker and whether the affected individual belongs to the group potentially subject to bias. Each extended Roy decision maker thus employs a simple cutoff rule: conditional on analyst observables, all individuals with expected potential outcomes on one side of the decision maker's cutoff are treated, and all those on the other side are untreated.[9] An unbiased decision maker in the extended Roy model is therefore

[8] The Roy model has been enormously influential in the analysis of economic decision making across a wide range of settings. Early applications include labor market participation (Gronau, 1974; Heckman, 1974), union membership (Lee, 1978), college attendance (Willis and Rosen, 1979), marriage (McElroy and Horney, 1981), domestic and international migration (Robinson and Tomes, 1982; Borjas, 1987), and occupational choice (Dolton et al., 1989).

[9] Exactly which potential outcomes are relevant for decision making will vary across empirical contexts. In pre-trial release, $Y_0 \equiv 0$ for all defendants, so $Y_1$ was the relevant potential outcome for an extended Roy bail judge. An extended Roy model of disability insurance examiners, conversely, would likely focus on $Y_0$, an applicant's potential for gainful employment in the absence of DI receipt. In other settings, like trial judges considering potential criminogenic effects of incarcerating first-time offenders, the causal effect of treatment $Y_1 - Y_0$ may be the relevant quantity.

bound to equalize the outcomes of marginal individuals across different groups, consistent with the outcome test. The extended Roy model also satisfies the instrument monotonicity condition required to econometrically identify the marginal contrasts of interest, since it is a special case of the separable decision model shown by Vytlacil (2002) to be equivalent to Imbens and Angrist (1994) monotonicity.

To decide whether an outcome test of bias based on the extended Roy model is appropriate in a given setting, detailed institutional knowledge about behavior, information sets, and measurement is essential. Nevertheless, a general set of guidelines applies: the analyst must be able to reasonably argue that decision makers harbor no other biases against any analyst-unobserved characteristics of the individuals they encounter; that the bias of interest is of constant magnitude across members of the unfavored group, invariant to other analyst-unobserved characteristics; that no errors in decision makers' outcome predictions, nor in the analyst's measurement of the outcome, correlate with any analyst-unobserved individual characteristics; and that no other decision considerations, beyond the single measured outcome of interest, vary across individuals along dimensions unobserved to the analyst. When making these arguments, it is important to keep in mind that these factors can invalidate the outcome test even if they are independent of the group membership the analyst is investigating as a potential target of bias. It is also useful to observe that the extended Roy model implies restrictions on the relationship between the treatment effects of interest and the probabilities of being treated, which can be tested in the data.

## References


American Bar Association (2007): *ABA Standards for Criminal Justice: Pretrial Release*, Washington, D.C.: American Bar Association, 3rd ed.

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

Arnold, D., W. Dobbie, and P. Hull (2020): "Measuring Racial Discrimination in Bail Decisions," Working paper.

Arnold, D., W. Dobbie, and C. S. Yang (2018): "Racial bias in bail decisions," *The Quarterly Journal of Economics*, 133, 1885–1932.

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

Ayres, I. and J. Waldfogel (1994): "A Market Test for Race Discrimination in Bail Setting," *Stanford Law Review*, 46, 987–1047.

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

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain," *Econometrica*, 80, 2369–2429.

Bertrand, M. and E. Duflo (2017): "Field Experiments on Discrimination," in *Handbook of Field Experiments*, ed. by A. Banerjee and E. Duflo, North-Holland, vol. 1, 309–393.

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

Björklund, A. and R. Moffitt (1987): "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *The Review of Economics and Statistics*, 69, 42–49.

Borjas, G. J. (1987): "Self-Selection and the Earnings of Immigrants," *The American Economic Review*, 77, 531–553.

Bouchard Ryan, E., M. A. Carranza, and R. W. Moffie (1977): "Reactions Toward Varying Degrees of Accentedness in the Speech of Spanish-English Bilinguals," *Language and Speech*, 20, 267–273.

Brinch, C. N., M. Mogstad, and M. Wiswall (2017): "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, 125, 985–1039.

Brock, W. A., J. Cooley, S. N. Durlauf, and S. Navarro (2012): "On the observational implications of taste-based discrimination in racial profiling," *Journal of Econometrics*, 166, 66–78.

Brownell, K. D., R. M. Puhl, M. B. Schwartz, and L. Rudd, eds. (2005): *Weight bias: Nature, consequences, and remedies.*, New York, NY, US: Guilford Publications.

Cozzarelli, C., A. V. Wilkinson, and M. J. Tagler (2001): "Attitudes Toward the Poor and Attributions for Poverty," *Journal of Social Issues*, 57, 207–227.

Dahl, G. B., A. R. Kostøl, and M. Mogstad (2014): "Family Welfare Cultures," *The Quarterly Journal of Economics*, 129, 1711–1752.

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.

Dobbie, W., J. Goldin, and C. S. Yang (2018): "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," *American Economic Review*, 108, 201–240.

Dobbie, W., A. Liberman, D. Paravisini, and V. Pathania (2020): "Measuring Bias in Consumer Lending," Working paper.

Dobbie, W. and J. Song (2015): "Debt Relief and Debtor Outcomes: Measuring the Effects of Consumer Bankruptcy Protection," *American Economic Review*, 105, 1272–1311.

Dolton, P. J., G. H. Makepeace, and W. van der Klaauw (1989): "Occupational Choice and Earnings Determination: The Role of Sample Selection and Non-Pecuniary Factors," *Oxford Economic Papers*, 41, 573–594.

Dovidio, J. F., M. Hewstone, P. Glick, and V. M. Esses (2010): *The SAGE Handbook of Prejudice, Stereotyping and Discrimination*, SAGE Publications.

Doyle, J. J. (2007): "Child Protection and Child Outcomes: Measuring the Effects of Foster Care," *American Economic Review*, 97, 1583–1610.

——— (2008): "Child Protection and Adult Crime: Using Investigator Assignment to Estimate Causal Effects of Foster Care," *Journal of Political Economy*, 116, 746–770.

Dunn, D. (2014): *The Social Psychology of Disability*, Academy of Rehabilitation Psychology Series, Oxford University Press.

Galasso, A. and M. Schankerman (2015): "Patents and Cumulative Innovation: Causal Evidence from the Courts," *The Quarterly Journal of Economics*, 130, 317–369.

Ghavami, N. and L. A. Peplau (2012): "An Intersectional Analysis of Gender and Ethnic Stereotypes: Testing Three Hypotheses," *Psychology of Women Quarterly*, 37, 113–127.

Gluszek, A. and J. F. Dovidio (2010): "The Way They Speak: A Social Psychological Perspective on the Stigma of Nonnative Accents in Communication," *Personality and Social Psychology Review*, 14, 214–237.

Gronau, R. (1974): "Wage Comparisons–A Selectivity Bias," *Journal of Political Economy*, 82, 1119–1143.

Hagiwara, N., D. A. Kashy, and J. Cesario (2012): "The independent effects of skin tone and facial features on Whites' affective reactions to Blacks," *Journal of Experimental Social Psychology*, 48, 892–898.

Hatfield, E. and S. Sprecher (1986): *Mirror, Mirror: The Importance of Looks in Everyday Life*, State University of New York Press.

Hebl, M. R. and J. M. Turchin (2005): "The Stigma of Obesity: What About Men?" *Basic and Applied Social Psychology*, 27, 267–275.

Heckman, J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679–694.

Heckman, J. J. (1998): "Detecting Discrimination," *Journal of Economic Perspectives*, 12, 101–116.

——— (2010): "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy," *Journal of Economic Literature*, 48, 356–398.

Heckman, J. J. and P. Siegelman (1993): "The Urban Institute Audit Studies: Their Methods and Findings," in *Clear and Convincing Evidence: Measurement of Discrimination in America*, ed. by M. Fix and R. Struyk, Urban Institute.

Heckman, J. J. and E. J. 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, vol. 6, 4875–5143.

Hester, N. and K. Gray (2018): "For Black men, being tall increases threat stereotyping and police stops," *Proceedings of the National Academy of Sciences*, 115, 2711 LP – 2715.

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Kang, S. K. and G. V. Bodenhausen (2015): "Multiple Identities in Social Perception and Interaction: Challenges and Opportunities," *Annual Review of Psychology*, 66, 547–574.

Kang, S. K. and A. L. Chasteen (2009): "Beyond the double-jeopardy hypothesis: Assessing emotion on the faces of multiply-categorizable targets of prejudice," *Journal of Experimental Social Psychology*, 45, 1281–1285.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018): "Human Decisions and Machine Predictions," *The Quarterly Journal of Economics*, 133, 237–293.

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

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

Koch, L. M., A. M. Gross, and R. Kolts (2001): "Attitudes Toward Black English and Code Switching," *Journal of Black Psychology*, 27, 29–42.

Lang, K. and J.-Y. K. Lehmann (2012): "Racial Discrimination in the Labor Market: Theory and Empirics," *Journal of Economic Literature*, 50, 959–1006.

Lang, K. and A. K.-L. Spitzer (2020): "Race Discrimination: An Economic Perspective," *Journal of Economic Perspectives*, 34, 68–89.

Lee, L.-F. (1978): "Unionism and Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables," *International Economic Review*, 19, 415–433.

Leslie, E. and N. G. Pope (2017): "The Unintended Impact of Pretrial Detention on Case Outcomes: Evidence from New York City Arraignments," *The Journal of Law and Economics*, 60, 529–557.

Livingston, R. W. and M. B. Brewer (2002): "What are we really priming? Cue-based versus category-based processing of facial stimuli," *Journal of Personality and Social Psychology*, 82, 5–18.

Maddox, K. B. (2004): "Perspectives on Racial Phenotypicality Bias," *Personality and Social Psychology Review*, 8, 383–401.

Maestas, N., K. J. Mullen, and A. Strand (2013): "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt," *American Economic Review*, 103, 1797–1829.

McElroy, M. B. and M. J. Horney (1981): "Nash-Bargained Household Decisions: Toward a Generalization of the Theory of Demand," *International Economic Review*, 22, 333–349.

Mogstad, M., A. Santos, and A. Torgovitsky (2018): "Using instrumental variables for inference about policy relevant treatment parameters," *Econometrica*, 86, 1589–1619.

Nelson, T. D. (2004): *Ageism: Stereotyping and Prejudice Against Older Persons*, MIT Press.

Persico, N. (2009): "Racial Profiling? Detecting Bias Using Statistical Evidence," *Annual Review of Economics*, 1, 229–254.

Robinson, C. and N. Tomes (1982): "Self-Selection and Interprovincial Migration in Canada," *The Canadian Journal of Economics / Revue canadienne d'Economique*, 15, 474–502.

Roy, A. D. (1951): "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3, 135–146.

Sampat, B. and H. L. Williams (2019): "How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome," *American Economic Review*, 109, 203–236.

Schneider, D. J. (2005): *The Psychology of Stereotyping*, Distinguished contributions in psychology, Guilford Publications.

Tucker, G. R. and W. E. Lambert (1969): "White and Negro Listeners' Reactions to Various American-English Dialects," *Social Forces*, 47, 463–468.

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341.

Wagner, U., O. Christ, and W. Heitmeyer (2010): "Anti-Immigration Bias," in *The SAGE Handbook of Prejudice, Stereotyping and Discrimination*, ed. by J. F. Dovidio, M. Hewstone, P. Glick, and V. M. Esses, SAGE Publications.

Willis, R. J. and S. Rosen (1979): "Education and Self-Selection," *Journal of Political Economy*, 87, S7–S36.

# Appendix

# A Additional Results on the Invalidity of the Outcome Test

In this section we present the proof of Theorem 3.1 and show that essentially the same theorem can be derived under significantly weaker conditions than those in Assumptions 3.1 and 3.2. Indeed, we do this by first presenting these alternative assumptions, then presenting the main theorem under these weaker assumptions in Theorem A.1, and then presenting the proofs of both Theorems as these are closely related.

## A.1 Results under Weaker Assumptions

Consider the following assumptions on the expected cost and benefit functions.

**Assumption A.1.** *For all* $(z,r) \in \mathcal{Z} \times \{w,b\}$ *there exists* $V^*_{z,r} \in \text{int}(\mathcal{V})$ *such that*

$$\Lambda(r, V^*_{z,r}) = \tau(z, r, V^*_{z,r}) \ , \tag{A.1}$$

$\Lambda(r,v) < \tau(z,r,v)$ *for all* $v < V^*_{z,r}$*, and* $\Lambda(r,v) > \tau(z,r,v)$ *for all* $v > V^*_{z,r}$*.*

**Assumption A.2.** *The range of the expected cost function* $\Lambda(r,v)$ *is the convex set* $\mathcal{I} \subseteq \mathbf{R}$*.*

**Assumption A.3.** *The expected benefit function* $\tau(z,r,\cdot)$ *satisfies the following conditions for all* $(z,r) \in \mathcal{Z} \times \{w,b\}$*:*

*(i) It maps* $\mathcal{V}$ *to the convex set* $\mathcal{I} \subseteq \mathbf{R}$

*(ii) For any* $V^*_{z,r}$ *as defined in* (A.1)*, there exists an open set* $\mathcal{V}^0_{z,r} \subseteq \mathcal{V}$ *such that*

$$\tau(z,r,v) > \tau(z,r,V^*_{z,r}) \quad \text{for all } v \in \mathcal{V}^0_{z,r} \ ,$$

*and* $\tau(z,r,\cdot)$ *is strictly decreasing over* $\mathcal{V}^0_{z,r}$*.*

Assumption A.1 is the same as Assumption 2.3 and is re-stated here for convenience. Assumption A.2 simply requires the expected cost function to take values in a convex set $\mathcal{I}$, which is allowed to be the entire real line or an interval like $[0,1]$. For example, when $\Lambda(r,v) = P\{Y_1 = 1 | R = r, V = v\}$ it follows that $\mathcal{I} = [0,1]$. Importantly, and as opposed to Assumption 3.1, note that Assumption A.2 does not require the function $\Lambda(r,\cdot)$ to be continuous or monotone. Assumption A.3(i) requires the expected benefit function to take values in $\mathcal{I}$ as well. This assumption is again only made for convenience and the same results could be derived when the range of the expected benefit and expected cost functions do not coincide at the expense

of some additional notation. Assumption A.3(ii) requires that $V^*_{z,r}$ is not a "maximum" of $\tau(z,r,\cdot)$. If the marginal defendant happens at the maximum value of $\tau(z,r,v)$, then it is impossible to find another intersection point with a higher value of $\tau(z,r,v)$ for the opposite race (and hence, a higher value of $\Lambda(r,v)$). So, there must be some place in the domain $\mathcal{V}$ where we can find higher values of the expected benefit $\tau(z,r,v)$. This is a mild assumption that, as opposed to Assumption 3.2, does not require this function to be continuous and monotone everywhere.

Under these assumptions we can prove the following result, where we use the notation $I_l \equiv \inf(\mathcal{I})$ and $I_u \equiv \sup(\mathcal{I})$.

**Theorem A.1.** *The following two results hold.*

*(i) Suppose that judge $z \in \mathcal{Z}$ is racially unbiased,*

$$I_l < \tau(z,r,v) = \tau(z,v) < I_u \text{ for all } v \in \mathcal{V} \ . \tag{A.2}$$

*Then, for any functions $\Lambda(b,v)$ and $\tau(z,v)$ jointly satisfying Assumptions A.1-A.3 there exists a function $\Lambda(w,v)$ satisfying Assumptions A.1-A.2 such that the marginal white defendant exhibits a higher misconduct probability than the marginal black defendant,*

$$\Lambda(w, V^*_{z,w}) > \Lambda(b, V^*_{z,b}) \ . \tag{A.3}$$

*(ii) Suppose that judge $z \in \mathcal{Z}$ discriminates against black defendants,*

$$\tau(z,w,v) > \tau(z,b,v) \text{ for all } v \in \mathcal{V} \ . \tag{A.4}$$

*Then, for any functions $\Lambda(w,v)$ and $\tau(z,w,v)$ jointly satisfying Assumptions A.1-A.3 there exist functions $\Lambda(b,v)$ and $\tau(z,b,v)$ jointly satisfying Assumptions A.1-A.3 such that the marginal black defendant exhibits a higher misconduct probability than the marginal white defendant,*

$$\Lambda(w, V^*_{z,w}) \leq \Lambda(b, V^*_{z,b}) \ . \tag{A.5}$$

*Proof.* Fix an arbitrary value of $z \in \mathcal{Z}$. Throughout the proof we make use of the following partition of $\mathcal{V}$ and the functions $H_r(v)$ and $G_r(v)$ defined below. Let $V^*_{z,r}$ be as in (A.1) and define the following partition of $\mathcal{V}$:

$$\mathcal{V}^{(-)}_{z,r} \equiv \{v \in \mathcal{V} : v \leq V^*_{z,r}\} \text{ and } \mathcal{V}^{(+)}_{z,r} \equiv \{v \in \mathcal{V} : v > V^*_{z,r}\} \ . \tag{A.6}$$

Let $H_r(v)$ be defined as follows:

$$H_r(v) = \{\text{continuous and weakly increasing function mapping } \mathcal{V}_{z,r}^{(-)} \text{ to } \mathcal{I} \\ \text{s.t. } H_r(v) < \tau(z,r,v)\ \forall v < V_{z,r}^* \text{ and } H_r(V_{z,r}^*) = \tau(z,r,V_{z,r}^*)\}\ . \tag{A.7}$$

Such a function always exists provided $\tau(z,r,v) > I_l$ for all $v \in \mathcal{V}_{z,r}^{(-)}$. In addition, let $G_r(v)$ be defined as follows:

$$G_r(v) = \{\text{continuous and strictly increasing function mapping } \mathcal{V}_{z,r}^{(+)} \text{ to } \mathcal{I} \\ \text{s.t. } G_r(v) > \tau(z,r,v)\ \forall v \in \mathcal{V}_{z,r}^{(+)}, \lim_{v \to V_{z,r}^*} G_r(v) = \tau(z,r,V_{z,r}^*), \lim_{v \to +\infty} G_r(v) = I_u\}\ . \tag{A.8}$$

Such a function always exists provided $\tau(z,r,v) < I_u$ for all $v \in \mathcal{V}_{z,r}^{(+)} \cup \{V_{z,r}^*\}$.

Part (i): We prove the statement by considering functions $\Lambda(b,v)$ and $\tau(z,v)$ satisfying AssumptionsA.1-A.3, and then constructing a function $\Lambda(w,v)$ satisfying Assumptions A.1-A.2 and the condition in (A.3). In addition, when constructing the function $\Lambda(w,v)$ we will do so by imposing continuity and weak monotonicity. This is unnecessary for sake of this result, but it simplifies the proofs of related results later.

Let $V_{z,b}^*$ be as in (A.1) and note that by Assumption A.3(ii) there exists an open set $\mathcal{V}_{z,b}^0 \in \mathcal{V}$ such that

$$\tau(z,v) > \tau(z,V_{z,b}^*) \ \text{ for all } v \in \mathcal{V}_{z,b}^0\ . \tag{A.9}$$

Pick an arbitrary point in $\mathcal{V}_{z,b}^0$ and denote it by $V_{z,w}^*$. Without loss of generality, assume $V_{z,w}^* < V_{z,b}^*$ (a symmetric argument applies otherwise). Next, define the function $\Lambda(w,v)$ by

$$\Lambda(w,v) = \begin{cases} H_w(v) & \text{for } v \in \mathcal{V}_{z,w}^{(-)} \\ G_w(v) & \text{for } v \in \mathcal{V}_{z,w}^{(+)} \end{cases}\ . \tag{A.10}$$

The function $H_w(v)$ is well defined since $\tau(z,v) > \Lambda(b,v) \geq I_l$ for all $v \in \mathcal{V}_{z,w}^{(-)}$ by Assumption A.1. The function $G_w(v)$ is well defined since $\tau(z,v) < I_u$ for all $v \in \mathcal{V}$ by (A.2). It then follows that Assumption A.1 immediately holds for $\tau(z,v)$ and $\Lambda(w,v)$ as in (A.10). In addition, $\Lambda(w,v)$ in (A.15) not only satisfies Assumption A.2 but is also continuous and weakly increasing. To conclude the proof, we note that (A.3) holds because

$$\Lambda(w,V_{z,w}^*) = \tau(z,V_{z,w}^*) > \tau(z,V_{z,b}^*) = \Lambda(b,V_{z,b}^*)\ ,$$

where the strict inequality follows from (A.9) and $V_{z,w}^* \in \mathcal{V}_z^0$.

Part (ii): We prove the statement by considering functions $\Lambda(w,v)$ and $\tau(z,w,v)$ satisfying Assumptions

A.1-A.3, and then constructing functions $\Lambda(b,v)$ and $\tau(z,b,v)$ satisfying Assumptions A.1-A.3, and the conditions in (A.4) and (A.5). In addition, when constructing the function $\Lambda(b,v)$ we will do so by imposing continuity and weak monotonicity. This is unnecessary for sake of this result, but it simplifies the proofs of related results later.

Let's start with $\tau(z,b,v)$. First note that $\tau(z,w,v)$ satisfies Assumptions A.1 and A.3 and that $\tau(z,w,v) > I_l$ for all $v \in \mathcal{V}$ must be true for (A.4) to happen. Next, let $V^*_{z,w}$ be as in (A.1) and note that Assumption A.3(ii) implies that there exists an open set $\mathcal{V}^0_{z,w} \in \mathcal{V}$ such that $\tau(z,w,v) > \tau(z,w,V^*_{z,w})$ for all $v \in \mathcal{V}^0_{z,w}$. Pick an arbitrary point in $\mathcal{V}^0_{z,w}$ and denote it by $V^*_{z,b}$. Define $\tilde{\epsilon}(v) \equiv \min\{\tilde{\epsilon}_1, \tilde{\epsilon}_2(v)\}$ where

$$\tilde{\epsilon}_1 \equiv \frac{1}{2}(\tau(z,w,V^*_{z,b}) - \tau(z,w,V^*_{z,w})) > 0 \tag{A.11}$$

$$\tilde{\epsilon}_2(v) \equiv \frac{1}{2}(\tau(z,w,v) - I_l) > 0 \ . \tag{A.12}$$

Using $\tilde{\epsilon}(v)$, we can then define $\tau(z,b,v)$ as

$$\tau(z,b,v) \equiv \tau(z,w,v) - \tilde{\epsilon}(v) \ , \tag{A.13}$$

for all $v \in \mathcal{V}$, and immediately satisfy (A.4). Note that $\tau(z,b,v) < I_u$ immediately follows by definition and $\tau(z,w,v) \leq I_u$. In addition, since $\tilde{\epsilon}(v) \leq \epsilon_2(v)$ and $\tau(z,w,v) > I_l$ for all $v \in \mathcal{V}$, it follows that $\tau(z,b,v) > I_l$ for all $v \in \mathcal{V}$ as well. Finally, consider $v, v' \in \mathcal{V}^0_{z,w}$ such that $v' > v$ and recall that $\tau(z,w,v) > \tau(z,w,v')$ by Assumption A.3(ii). Note that

$$\tilde{\epsilon}(v) - \tilde{\epsilon}(v') \leq \frac{1}{2}(\tau(z,w,v) - \tau(z,w,v')) < \tau(z,w,v) - \tau(z,w,v') \ , \tag{A.14}$$

so that by (A.13) $\tau(z,b,v) > \tau(z,b,v')$ and $\tau(z,b,v)$ is strictly decreasing over $\mathcal{V}^0_{z,w}$. Since $V^*_{z,b} \in \mathcal{V}^0_{z,w}$, we can define the open set $\mathcal{V}^0_{z,b} \equiv \mathcal{V}^0_{z,w} \cap \{v \in \mathcal{V} : v < V^*_{z,b}\}$ over which Assumption A.3 holds for $\tau(z,b,v)$.

Let's now consider the function $\Lambda(b,v)$. Partition $\mathcal{V}$ according to (A.6) and define the function $\Lambda(b,v)$ as

$$\Lambda(b,v) = \begin{cases} H_b(v) & \text{for } v \in \mathcal{V}^{(-)}_{z,b} \\ G_b(v) & \text{for } v \in \mathcal{V}^{(+)}_{z,b} \end{cases} \tag{A.15}$$

where $H_b(v)$ is as in (A.7) and $G_b(v)$ is as in (A.8). The function $H_b(v)$ is well defined since $\tau(z,b,v) > I_l$ for all $v \in \mathcal{V}$ by construction. The function $G_b(v)$ is also well defined since $\tau(z,b,v) < I_u$ for all $v \in \mathcal{V}$ by construction as well. It then follows that Assumption A.1 immediately holds for $\tau(z,b,v)$ as in (A.13) and $\Lambda(b,v)$ as in (A.15). In addition, $\Lambda(b,v)$ in (A.15) not only satisfies Assumption A.2 but is also continuous

and weakly increasing. To conclude the proof, we note that (A.5) holds because

$$\Lambda(b, V^*_{z,b}) = \tau(z, b, V^*_{z,b}) \geq \tau(z, w, V^*_{z,b}) - \tilde{\epsilon}_1 > \tau(z, w, V^*_{z,w}) = \Lambda(w, V^*_{z,w}) \ , \tag{A.16}$$

where the weak inequality follows from $-\tilde{\epsilon}(v) \geq -\tilde{\epsilon}_1$ and the strict inequality follows from the definition of $\tilde{\epsilon}_1$ in (A.11). This concludes the proof of part (ii). It is worth noting that under the additional restriction that $\Lambda(r, v) = \Lambda(v)$ for all $r \in \{b, w\}$ it is possible to construct the function $\Lambda(\cdot)$ in a way such that $\Lambda(V^*_{z,b}) = \Lambda(V^*_{z,w})$. We omit the proof of this statement as it is similar to the one above, but present the graphical intuition in Figure A.3a. This completes the proof. ■

**Remark A.1.** *The condition in* (A.2) *requires the function* $\tau(z, v)$ *to take values in* $(I_l, I_u)$*. This requirement is needed to avoid highly discontinuous functions that suddenly take values equal to* $I_l$ *or* $I_u$ *in the interior of* $\mathcal{V}$*, therefore preventing finding expected cost functions that could be strictly above or below such functions (which is required for Assumption 2.3). However, if additional smoothness conditions on the function* $\tau(z, v)$ *are imposed, for example as in Assumption 3.2, this requirement can be removed. This is indeed the case in Theorem 3.1.*

In the next section we prove Theorem 3.1 by exploiting Lemma A.1 below, which establishes that Assumptions 3.1 and 3.2 imply Assumptions A.2 and A.3. Give this lemma, the proof proceeds by showing that the results in Theorem 3.1 essentially follow from the proof of Theorem A.1.

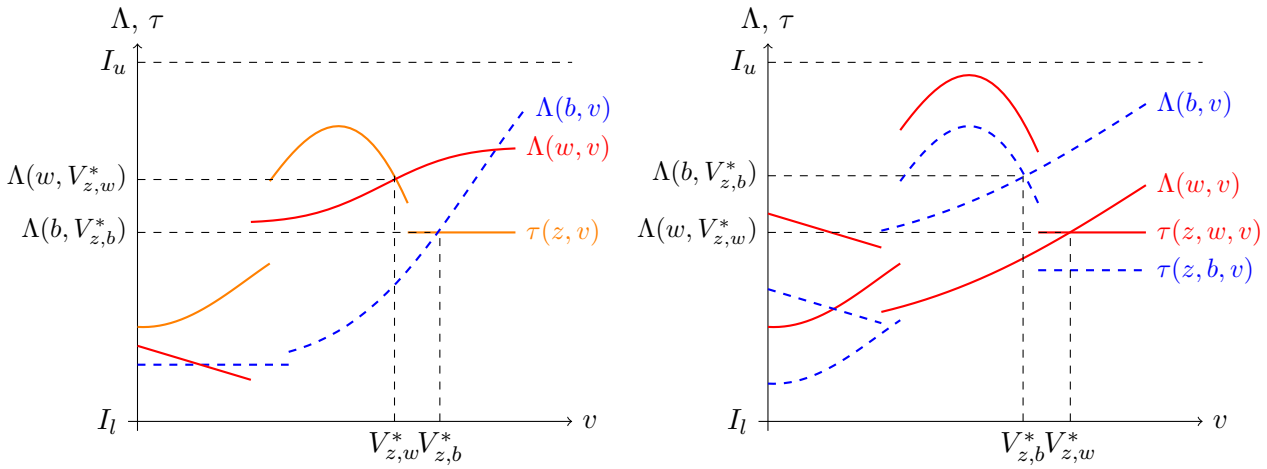
**Lemma A.1.** *Let Assumption 2.3 hold. Then Assumptions 3.1 and 3.2 imply Assumptions A.2 and A.3.*

*Proof.* Assumption 3.1 trivially implies A.2 as the continuous image of a connected set (i.e., $\mathcal{V}$) is connected, and so $\mathcal{I}$ is an interval in $\mathbf{R}$ (and thus convex). Similarly, Assumption A.3 trivially follows from 3.2. ■

## A.2 Proof of Theorem 3.1

The proof of Theorem 3.1 follows immediately from the proof of Theorem A.1 after noticing that the function $\tau(z, b, v)$ defined in (A.13) satisfies Assumption 3.2 by construction, as $\tilde{\epsilon}(v)$ is continuous and, by (A.14), $\tau(z, b, v)$ is weakly decreasing for all $v \in \mathcal{V}$ and strictly decreasing whenever $\tau(z, w, v)$ is strictly decreasing. In addition, the functions $\Lambda(b, v)$ in (A.15) and $\Lambda(w, v)$ in (A.10) satisfy Assumption 3.1 by construction as well.

Finally, to see why $\tau(z, v) \in (I_l, I_u)$ in (A.2) is not needed in (6), note that Assumption 3.2 guarantees that $V^*_{z,w} < V^*_{z,b}$ in the proof of Part (i) of the Theorem. Given this, $\tau(z, v) > I_l$ is not needed as this requirement was only needed in cases where $V^*_{z,w} > V^*_{z,b}$. Finally, for $v \in \mathcal{V}^{(+)}_{z,w}$ Assumption 3.2 implies that $\tau(z, v) < \tau(z, V^*_{z,w}) < \tau(z, V^*_{z,w} - \epsilon) \leq I_u$ for some $\epsilon > 0$, since $V^*_{z,w} \in \text{int}(\mathcal{V})$ by Assumption 2.3. It follows that the function $G_w(v)$ is well-defined and this completes the proof.




(a) Generalization of Part (i)

(b) Generalization of Part (ii)

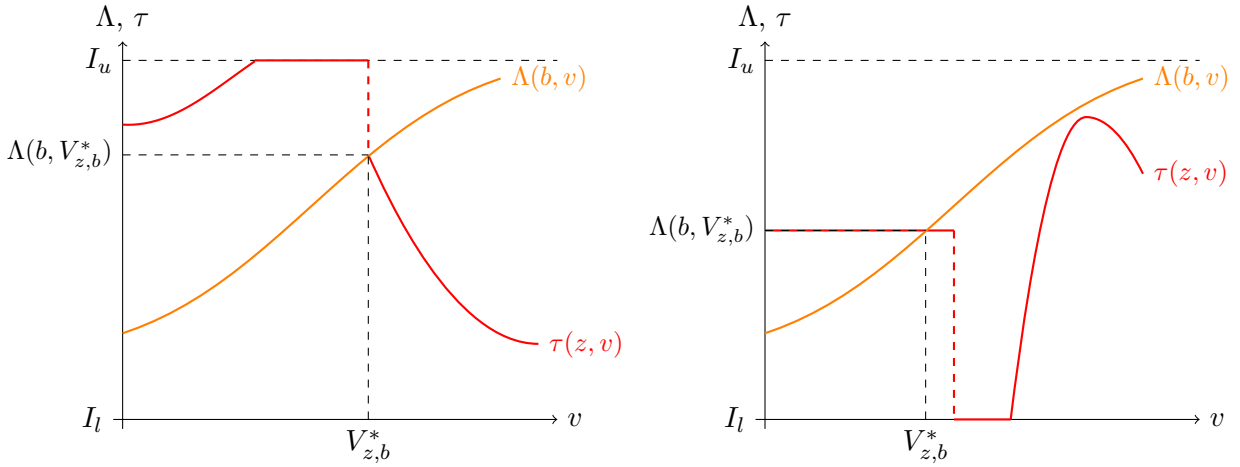
Figure A.1: Intuition behind Theorem A.1 under weaker assumptions

## A.3 Some Graphical Intuition

Figure A.1 illustrates the result in Theorem A.1 in the same way that Figure 1 illustrated the result in Theorem 3.1. In particular, the functions in Figure A.1 explicitly violate the assumptions required by Theorem 3.1, i.e. continuity and monotonicity, to highlight the differences between Theorems A.1 and 3.1.

Figure A.2a illustrates why (A.2) requires $\tau(z, v) < I_u$. If $\tau(z, v) = I_u$ and the intersection of $\tau(z, v)$ and $\Lambda(b, v)$ is at a point of discontinuity (or to the right of that), then any other function $\Lambda(\cdot)$ that intersects with $\tau(z, v)$ to the left of $V^*_{z,b}$ will suffer from two problems. First, it will not be able to be strictly above $\tau(\cdot)$ after such an intersection and, second, it will necessarily intersect with $\tau(z, v)$ at multiple points. This violates Assumption A.1. In turn, Figure A.2b illustrates why (A.2) requires $\tau(z, v) > I_l$. In the figure, any function $\Lambda(\cdot)$ that intersects with $\tau(z, v)$ at a higher point than $\Lambda(b, V^*_{z,b})$ will do so to the right of $V^*_{z,b}$ and will immediately violate Assumption A.1. Both of these situations arise when the function $\tau(z, v)$ is non-monotonic and this is why we do not need these conditions in Theorem 3.1.

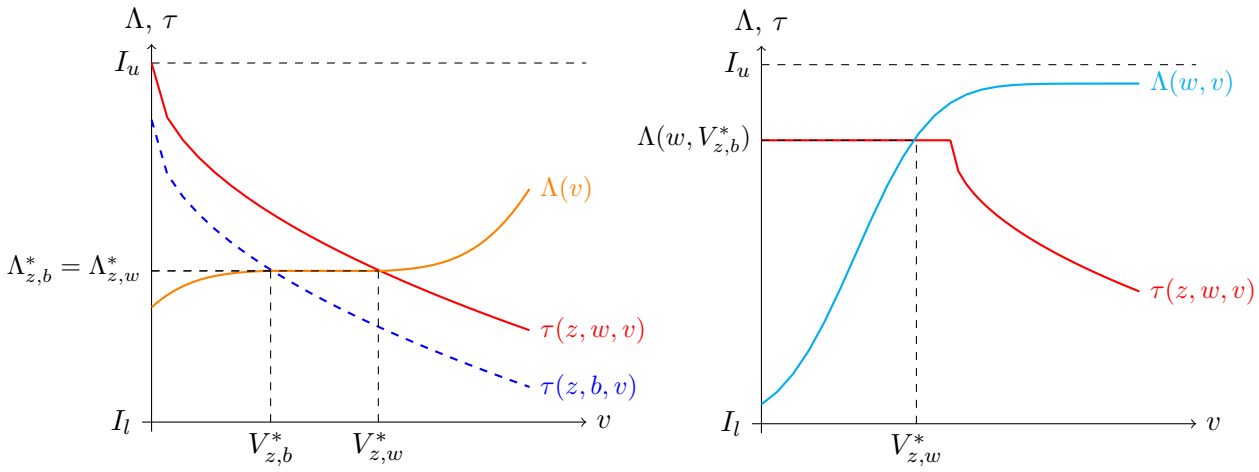
Figure A.3a illustrates a situation where the function $\Lambda(\cdot)$ does not vary by race and yet (9) (or (A.5)) happens with equality. Finally, Figure A.3b shows that if the marginal white defendant delivers the highest value of the function $\tau(z, w, v)$, then it is not possible to construct a function $\tau(z, b, v)$ that lies below $\tau(z, w, v)$ (as required by condition (A.4) or (8)) and delivers the situation in (9) (or (A.5)).




(a) Problem with $\tau(z, v) = I_u$

(b) Problem with $\tau(z, v) = I_l$

Figure A.2: Graphical intuition of why the condition $\tau(z, v) \in (I_l, I_u)$ is required in (A.2).




(a) Equation (9) (or (A.5)) with equality

(b) Marginal defendant at the maximum

Figure A.3: Additional graphical intuition for Theorems 3.1 and A.1.

## A.4 Proof of the Lemmas in Section 4

*Proof of Lemma 4.1.* Suppose that judge $z \in \mathcal{Z}$ is racially unbiased, i.e., $\tau(z, r, v) = \tau(z, v)$ for all $v \in \mathcal{V}$. By Assumption 2.3 and the fact that $\Lambda(r, v) = \Lambda(v)$, we have that

$$\tau(z, V^*_{z,w}) = \Lambda(w, V^*_{z,w}) = \Lambda(V^*_{z,w}) \text{ and } \tau(z, V^*_{z,b}) = \Lambda(b, V^*_{z,b}) = \Lambda(V^*_{z,b}) \ .$$

By Assumption 2.3, it must be the case that $V^*_{z,w} = V^*_{z,b}$ and so the outcome test is logically valid according to Definition 3.1.

Suppose that judge $z \in \mathcal{Z}$ discriminates against black defendants, i.e.

$$\tau(z, w, v) > \tau(z, b, v) \text{ for all } v \in \mathcal{V} \ . \tag{A.17}$$

The other direction involves a symmetric argument so we only present this case. Now split the argument in three cases. First, suppose that $V^*_{z,w} = V^*_{z,b}$ and consider the following argument,

$$\Lambda(V^*_{z,w}) = \tau(z, w, V^*_{z,w}) > \tau(z, b, V^*_{z,w}) > \Lambda(V^*_{z,w}) \ , \tag{A.18}$$

where the equalities follow from Assumption 2.3 and the inequality follows from (A.17). This is a contradiction. Second, suppose that $V^*_{z,w} < V^*_{z,b}$ and consider the following argument,

$$\Lambda(V^*_{z,w}) = \tau(z, w, V^*_{z,w}) > \tau(z, b, V^*_{z,w}) > \Lambda(V^*_{z,w}) \ , \tag{A.19}$$

where the equality follows from Assumption 2.3, the first inequality follows from (A.17), and the last inequality follows from Assumption 2.3 and the fact that $V^*_{z,w} < V^*_{z,b}$. This again leads to a contradiction. Finally, suppose then that $V^*_{z,w} > V^*_{z,b}$. Since the function $\Lambda(\cdot)$ is assumed to be strictly increasing, it follows that $\Lambda(V^*_{z,w}) > \Lambda(V^*_{z,b})$. It follows that the outcome test is logically valid according to Definition 3.1. ■

*Proof of Lemma 4.2.* Suppose that judge $z \in \mathcal{Z}$ is racially unbiased, i.e., $\tau(z, r) = \tau(z)$. By Assumption 2.3 we have that

$$\Lambda(w, V^*_{z,w}) = \tau(z) = \Lambda(b, V^*_{z,b}) \ ,$$

and so the outcome test is logically valid according to Definition 3.1.

Suppose that judge $z \in \mathcal{Z}$ discriminates against black defendants, i.e., $\tau(z, w) > \tau(z, b)$. The other direction involves a symmetric argument so we only present this case. By Assumption 2.3 we have that

$$\Lambda(w, V^*_{z,w}) = \tau(z, w) > \tau(z, b) = \Lambda(b, V^*_{z,b}) \ ,$$

and so the outcome test is logically valid according to Definition 3.1. ■