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, Revised June 2023

We thank the editor, three anonymous referees, Chuck Manski, and Xavier D'Haultfoeuille for comments. We are grateful for financial support from the Robert H. Topel Faculty Research Fund at the University of Chicago Booth School of Business. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

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, Revised June 2023
JEL No. C21,C26,C51,J15,J16,J71,K14,K42

ABSTRACT

The decisions of judges, lenders, journal editors, and other gatekeepers often lead to significant disparities 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, 1993) proposed an outcome test of bias based on differences in post-decision outcomes across groups, inspiring a large and growing empirical literature. The goal of our paper is to offer a methodological blueprint for empirical work that seeks to use outcome tests to detect bias. We show that models of decision making underpinning outcome tests can be usefully recast as Roy models, since heterogeneous potential outcomes enter directly into the decision maker's choice equation. Different members of the Roy model family, however, are distinguished by the tightness of the link between potential outcomes and decisions. We show that these distinctions have important implications for defining bias, deriving logically valid outcome tests of such bias, and identifying the marginal outcomes that the test requires.

Ivan A. Canay
3423 Kellogg Global Hub
Department of Economics
Northwestern University
2211 Campus Drive
Evanston, IL 60208
iacanay@northwestern.edu

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

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

1 Introduction

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

To answer this question, it is necessary to first define what it means for a decision to be unbiased. This requires specifying what an *unbiased* decision maker in a particular setting is supposed to be optimizing, what constraints they face, and what they should know at the time they make their decision. 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 of bias proposed by Becker (1957, 1993). Becker (1957)'s key insight is that a biased decision maker "must act *as if* he were willing to pay something" to exercise bias. Frequently used benchmark tests, which compare decision *rates* across groups, need not be informative about such bias, as differences in decision rates may simply reflect group differences in relevant individual characteristics that are unobserved by the analyst. Observing the downstream *outcomes* that result from a decision, on the other hand, could potentially be more informative about bias, as differences in these outcomes across groups may reveal the decision maker's "willingness to pay" to favor some groups over others. Becker (1993) offers an intuitive example in the context of mortgage lending: if a bank's loans to approved black applicants tend to generate higher profits than its loans to approved white applicants, this may (under certain assumptions) indicate racially biased lending decisions that hold black applicants to a stricter standard beyond what unbiased profit maximization would imply.

In this paper, we carefully examine what researchers can learn about bias in decision making from outcome tests, which compare post-decision outcomes across different groups. The chief goal is to offer a methodological blueprint for empirical work that seeks to use outcome tests to detect

bias. Much of the discrimination literature since Becker (1957) has focused on identification issues that arise when individuals differ along dimensions beyond the group membership of interest.[1] In outcome tests, such heterogeneity is known to render *average* outcomes across groups generally uninformative about the *marginal* cases that may reveal a decision maker's differential standards, as average comparisons confound any differences at the margin with distributional differences away from the margin (Peterson, 1981; Heckman, 1998; Ayres, 2002).[2] Innovative solutions to such "inframarginality bias" include equilibrium models in which the average individual endogenously behaves like, and thus identifies, the marginal (Knowles et al., 2001); partial identification approaches that derive testable implications of bias in comparisons of decision rates and average outcomes across different decision makers (Anwar and Fang, 2006; Alesina and La Ferrara, 2014); and instrumental variable strategies that use exogenous assignment of cases to decision makers of varying leniency to directly identify the outcomes of marginal cases (Arnold et al., 2018; Dobbie et al., 2021).

While addressing inframarginality bias is an essential input into a viable outcome test, we initially take a step back to ask a more fundamental question: under what conditions are the outcomes of marginal cases, even if known perfectly, actually informative about decision maker bias? To answer this question, we first lay out a modeling framework in Section 2 that describes decision maker behavior and sets the stage for defining and detecting bias in decision making. We focus on settings where a standalone decision maker makes a binary decision by maximizing the expected net benefit of that decision.[3] For concreteness, we ground our exposition in the context of racial bias in pre-trial release, where a bail judge decides whether to release or detain defendants of different races who are awaiting trial, as this setting features prominently in the recent outcome test literature (e.g. Kleinberg et al., 2018; Arnold et al., 2018, 2022). Our analysis generalizes, however, to a wide range of empirical settings and dimensions of bias, as we discuss throughout the paper.

We show that models of decision making underpinning outcome tests can be usefully recast as

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

[2] See also Brock et al. (2012), who study the mapping from *average* outcome comparisons to racial bias, and thus focus on the inframarginality complications that arise when distributions of unobservable characteristics differ by race. In contrast, we study the distinct and more fundamental mapping from *marginal* outcomes to bias; our main results are therefore insensitive to racial differences in the distribution of unobservables among inframarginal individuals, which is a central concern in Brock et al. (2012) and the entire literature on average-based outcome tests.

[3] For examples of outcome tests in settings with continuous, rather than binary, decision variables, see Ayres and Waldfogel (1994), Anwar and Fang (2012), Anwar and Fang (2015), and Mechoulan and Sahuguet (2015).

Roy models, since heterogeneous potential outcomes of the individuals subject to the decision—in the case of pre-trial release, the risk of flight or additional crimes ("misconduct") a defendant may commit before trial—enter directly into the decision maker's choice equation. Indeed, the intuitive motivation for the outcome test rests on an implicit assumption of some link between potential outcomes and decisions; otherwise, there would be no reason to expect outcomes to be informative about (bias in) decision making. Different members of the Roy model family, however, are distinguished by the tightness of the link between potential outcomes and decisions. At one extreme of the Roy spectrum, decision makers in the original Roy (1951) model simply choose the option with the most advantageous potential outcome, forging a very tight link between potential outcomes and decisions. At the other end of the spectrum is the Generalized Roy Model (Björklund and Moffitt, 1987; Heckman and Vytlacil, 2005), in which potential outcomes are allowed to, but not required to, influence decisions, as other factors unobservable to the econometrician are allowed to determine choices in unrestricted ways.

For the purposes of the outcome test, we show that a constructive middle ground between these two extremes is the Extended Roy Model (Heckman and Vytlacil, 2007; D'Haultfœuille and Maurel, 2013). In this model, a decision maker acts as if she follows a cutoff rule, comparing the expected effect of her decision on the outcome against a cutoff value. This cutoff value is allowed to vary across different decision makers (e.g. judges with different levels of leniency), and potentially across the group membership of interest in the outcome test (e.g. the race of the defendant), but not across any dimensions of heterogeneity unobserved by the econometrician, which is the key restriction of the Extended Roy Model that distinguishes it from the Generalized Roy Model. We show that these two members of the Roy model family implicitly underlie nearly all of the existing outcome test literature, and we show how their distinction has important implications not only for decision maker behavior, but also for the task of defining bias (Section 3), for the logical validity of the outcome test as an indicator of such bias (Section 4), and for the econometric identification of the marginal outcomes that the test requires (Section 5).

In Section 3, we use the models laid out in Section 2 to define bias and its absence, and discuss how our definitions relate to and nest notions of bias from the previous literature. Our framework explicitly distinguishes taste-based bias from other unrelated factors that can end up being empirically indistinguishable from it, including errors in the decision maker's outcome predictions, errors

in the econometrician's measurement or specification of the outcome variable, and other decision maker objectives beyond the measured outcome. Our various definitions of bias also show how the task of defining bias necessarily interacts with the task of choosing a decision model, revealing a tradeoff between simpler definitions of bias versus richer models of decision making.

In Section 4, we formally describe the outcome test, which aims to detect decision maker bias by comparing the post-decision outcomes of marginal individuals from different groups.[4] We then present two sets of results on the logical validity of this test. The first set of results show that the test is logically invalid in the context of the Generalized Roy Model. We show this by generating broad classes of counterexamples which prove that the outcome test may conclude that a decision maker is unbiased even if she is biased, or conclude bias even if the decision maker is unbiased, or even conclude bias against one group when the decision maker is actually biased against another. These logical invalidations arise because the Generalized Roy Model is not sufficiently restrictive to deliver an optimality condition in which an unbiased decision maker must equalize marginal outcomes across groups, since other factors unobserved to the econometrician can also influence decisions. One perhaps surprising lesson from our analysis is that these unobserved factors can invalidate the outcome test even if they are completely independent of the group membership potentially subject to bias, going beyond typical concerns in the discrimination literature about correlated confounders.

These findings raise the question of what additional restrictions are needed for the decision model to generate an optimality condition in which an unbiased decision maker equalizes marginal outcomes across groups, and thereby enable a logically valid outcome test. Our second set of results in Section 4 establish logical validity of the outcome test when restricting decision maker behavior to the Extended Roy Model, as this restriction shuts down the confounding channels present in the Generalized Roy Model. Imposing the Extended Roy Model has substantive implications for admissible decision maker behavior, so careful consideration must be paid to the credibility of those implications in any given empirical environment. We also discuss alternative ways to generate logically valid outcome tests, including changing the parameter of interest to alternative notions of bias or discrimination (e.g., Arnold et al., 2020, 2022), or restricting the data generating process

[4]Our analysis applies to examining not only whether bias exists, but also its magnitude, as well as heterogeneity in bias across different decision makers, including groups of decision makers. This has implications, for example, for the literature on in-group bias (e.g., Shayo and Zussman, 2011; Anwar et al., 2012; Knepper, 2018; Ash et al., 2022).

to eliminate any statistical relationship between potential outcomes and the group membership of interest.

Since all of our results on the logical validity of outcome tests in Section 4 assume that the outcomes of marginal members of each group are perfectly known, these results are distinct from, and prior to, challenges that arise in attempting to identify marginal outcomes from observed data, which we turn to in Section 5. Since the analyst rarely observes which particular individuals are marginal, we begin in Section 5.1 by considering the best-case econometric scenario of local instrumental variation across continuously distributed and exogenously assigned decision makers. We show that under the Extended Roy Model of decision maker behavior, the local instrumental variables estimand of Heckman and Vytlacil (2005) successfully recovers the outcome test's parameters of interest as group-specific marginal treatment effects, enabling both a logically valid and econometrically viable outcome test.[5] We also show that the Extended Roy Model has clear testable implications, establishing necessary conditions for its suitability in any given empirical setting.

In Section 5.2, we examine whether these results extend to other data environments and identification approaches. We first consider settings where point-identifying marginal outcomes is challenged by discretely distributed decision makers. We then study the frequent case where no valid instrumental variation across decision makers is available, and the analyst attempts to compare *average* outcomes across groups for a given decision maker or group of decision makers. We highlight the additional challenges to identification that arise in these environments and discuss potential solutions.

2 Framework

In this section, we lay out a modeling framework that describes decision maker behavior and sets the stage for defining and detecting bias in decision making. We focus on settings where a single agent makes a binary decision through a cost-benefit analysis. For concreteness, we ground our exposition in the context of racial bias in pre-trial release, where a bail judge decides whether to

[5] In contrast, we show in Appendix B that the Generalized Roy Model not only fails to deliver a logically valid test, but also does not permit identification of the marginal outcomes of interest without additional restrictions. Importantly, imposing such restrictions to ensure identification of MTEs does not address the issues of logical failure that we identify in Section 4. Without a sufficiently restrictive decision model, an analyst may therefore successfully identify marginal outcome differences across groups that are nevertheless uninformative about decision maker bias.

release or detain defendants of different races who are awaiting trial. Our analysis generalizes, however, to a wide range of empirical settings and dimensions of bias, as we discuss throughout.

2.1 Setup and Notation

The random vector $(Z, D, R, V, Y_1, Y_0, C)$ captures the relevant variables in the decision model. Each draw from this vector represents a case exogenously assigned to a decision maker Z who makes a binary decision D regarding an individual characterized by (R, V, Y_1, Y_0, C), defined below. Depending on the setting, the set of decision makers $\mathcal{Z}$ could be viewed as a continuum in $\mathbf{R}$ or a discrete set with few or many elements. The cardinality of $\mathcal{Z}$ does not affect our results on logical validity of the outcome test, which only concerns the behavior of a given decision maker z arbitrarily chosen from $\mathcal{Z}$. In Section 5, we discuss econometric viability of the outcome test using variation across both continuous and discrete Z.

In the pre-trial release environment, $D = 1$ if the bail judge decides to release the defendant prior to trial, and $D = 0$ if the defendant is detained. The judge observes the race R of each defendant, with $R = w$ denoting a white defendant and $R = b$ denoting a black defendant, as well as non-race characteristics V like criminal history, employment status, and family structure. We use lower case letters (r, v) to refer to the characteristics of a specific defendant.

The defendant's measured outcome Y, which is used to conduct the outcome test, takes values in $\mathcal{Y} \subseteq \mathbf{R}$, with potential outcome Y_1 occurring if the decision is $D = 1$ and Y_0 occurring if $D = 0$. Throughout, we maintain the exclusion restriction that a judge only affects defendant outcomes through her release decision, not through any direct effects on potential outcomes. Existing empirical work on bias in the pre-trial setting typically restricts (Y_1, Y_0) in two additional ways. First, Y is binarized, with Y_1 indicating whether the defendant would commit any pre-trial misconduct if released. Second, detention is assumed to prevent any misconduct from occurring, such that $Y_0 \equiv 0$ for all defendants. We do not need to impose these setting-specific restrictions in our framework, since they are immaterial for our results on logical validity and econometric viability of outcome tests. Instead, we work more generally with $\Delta \equiv Y_1 - Y_0$, the causal effect of release on a given defendant's measured misconduct outcome. This generality can be useful in other settings where Y is not binary or Y_0 is not equal to zero.[6]

[6] Under the restrictions typically imposed in the pre-trial release literature, $\Delta = Y_1 \in \{0, 1\}$. In other settings,

To capture bias, we denote by $\beta(z, r, v)$ the taste for discrimination parameter that judge z has for a defendant with characteristics (r, v). We devote Section 3 to formal definitions of bias using this parameter and discuss how they relate to previous definitions from the literature.

Finally, C denotes the cost of detaining a given defendant prior to trial. This cost captures both defendant-invariant considerations like local jail capacity and judge z's general level of leniency, as well as any defendant-specific costs of detention the judge might consider, including job loss, family instability, increased likelihood of conviction due to reduced bargaining power, and long-run criminogenic effects (Leslie and Pope, 2017; Dobbie et al., 2018). In general, C serves to collect other decision maker considerations beyond bias β and the specific outcome Y used to conduct the outcome test, i.e. "omitted payoffs" in the language of Kleinberg et al. (2018).

Remark 2.1. We purposefully defer discussing which of these variables are observable and unobservable to the econometrician until turning our attention to identification in Section 5. In doing so, we emphasize that the results we derive on logical validity of the outcome test are distinct from issues about econometric identification.

Remark 2.2. We also need not be concerned at this point whether information sets and beliefs may differ across judges. This is because logical validity of the outcome test is solely concerned with the mapping between the bias of any given judge and the outcomes of that specific judge's marginal black and white defendants.

2.2 Decision Model

Judge z makes a decision by minimizing expected cost, given her information at the time of the bail hearing. With the notation set out above, the problem solved by judge z can be written as

$$\min_{d \in \{0,1\}} \mathcal{E}_z[Y_d^* + (1-d)(C + \beta(z, R, V)) \mid R = r, V = v] \ , \tag{1}$$

where $\mathcal{E}_z[\cdot|R = r, V = v]$ denotes the judge's subjective conditional expectation, and Y_d^* is the potential outcome that judge z takes into account to make a decision, which potentially differs from the measured Y_d as we discuss below.

decision makers may focus only on Y_0, like disability insurance examiners assessing an applicant's potential for gainful employment in the absence of DI receipt D. Our framework incorporates all of these variations by defining Δ as whatever level or contrast of potential outcomes is relevant for the decision maker.

Judge z releases a defendant with characteristics (r, v) whenever the perceived misconduct cost of releasing the defendant, $\mathcal{E}_z[Y_1^* - Y_0^* | R = r, V = v]$, does not exceed the perceived cost of detaining the defendant $\mathcal{E}_z[C|R = r, V = v] \equiv c(z, r, v)$, where the cost of detention is shifted by the taste for discrimination parameter $\beta(z, r, v)$:

$$D_z = I\left\{\mathcal{E}_z[Y_1^* - Y_0^* | R = r, V = v] \leq c(z, r, v) + \beta(z, r, v)\right\} \ . \tag{2}$$

In this way, a high value of $\beta(z, r, v)$ means that judge z exacerbates the cost of detaining a defendant with characteristics (r, v).[7]

This model captures important features of bail judge decision making, and more generally, makes precise three key elements that describe how a given decision maker is assumed to behave in a given setting. The first key element is the specification of the *objectives* of the decision maker. In most jurisdictions, bail judges are instructed to detain defendants who "pose a substantial risk of flight, or threat to the safety of the community, victims or witnesses or to the integrity of the justice process," while simultaneously minding the costs of such detention: "Deprivation of liberty pending trial is harsh and oppressive, subjects defendants to economic and psychological hardship, interferes with their ability to defend themselves, and, in many instances, deprives their families of support" (American Bar Association, 2007). The model captures this tradeoff between misconduct Y_d^* and detention costs C in the objective function, while also allowing for the possibility of biases β that tilt the scales of this tradeoff to unfairly favor of some types of defendants over others.

The second key element is the specification of the *information* that the decision maker uses to assess these objectives. Bail judges are typically given broad discretion to use "any facts justifying a concern that a defendant will present a serious risk of flight or of obstruction, or of danger to the community or the safety of any person" if released (American Bar Association, 2007). The model naturally captures this broad use of "any facts" through the conditioning set $(R = r, V = v)$, which by definition includes all of the defendant and case characteristics that the judge is able to observe at the time of the hearing.

The final key element is the specification of how the decision maker forms *beliefs* about the

[7] In our formulation, the taste for discrimination parameter $\beta(z, r, v)$ enters additively in the objective function in (1). This is the formulation adopted, for example, by Knowles et al. (2001) and Anwar and Fang (2006). An alternative formulation would introduce $\beta(z, r, v)$ multiplicatively, similar to Persico (2009). This modeling choice does not affect the substantive points we make in this paper.

objectives given the information she observes. The term $\mathcal{E}_z[Y_1^* - Y_0^*|R = r, V = v]$ denotes the expected misconduct cost of release given the beliefs of judge z, with the notation explicitly allowing for this subjective expectation to deviate from $E[Y_1 - Y_0|R = r, V = v]$, which we hereafter write more succinctly as $E[\Delta|r, v]$. To make this distinction clear, we explicitly define such deviations as

$$\lambda(z, r, v) \equiv E[\Delta|r, v] - \mathcal{E}_z[Y_1^* - Y_0^*|r, v] \ . \tag{3}$$

The function $\lambda(z, r, v)$ captures a variety of factors that may drive a wedge between the mean effect of release on misconduct among defendants with characteristics (r, v) and judge z's prediction of that effect.[8] First, $\lambda(z, r, v)$ captures any systematic measurement error in the outcome by allowing for $Y_d^* \neq Y_d$. Defendants with more criminal experience, for example, may be more likely to commit additional crimes while awaiting trial that go unsolved by law enforcement, driving a wedge between measured misconduct Y and the actual misconduct the judge tries to predict, Y^*. Similarly, $\lambda(z, r, v)$ captures any specification error in the analyst's definition of Y, e.g., when Y is a binarized measure of any pre-trial misconduct but judge z considers a more continuous measure of misconduct severity. Lastly, even if $Y_d^* = Y_d$, judge z may make prediction errors that vary systematically with defendant characteristics, i.e. $E[\Delta|r, v] \neq \mathcal{E}_z[\Delta|r, v]$. The results in Kleinberg et al. (2018), for example, suggest that bail judges overestimate misconduct among defendants who face a serious current charge but have little criminal history, and conversely underestimate misconduct among defendants who face a minor current charge but have serious prior convictions.

Remark 2.3. Many seminal papers in the outcome test literature, including Knowles et al. (2001), Anwar and Fang (2006), and Persico (2009), implicitly or explicitly assume $\lambda(z, r, v)$ to be zero. In both Knowles et al. (2001) and Anwar and Fang (2006), for example, a police officer's expected benefit of stopping a driver equals the true probability that the driver is carrying contraband. Likewise, in the framework of Persico (2009), decision makers know the true expected profits of every action. More recent studies of racial bias in pre-trial release, e.g., Arnold et al. (2018), Arnold et al. (2022), and Hull (2021), discuss judge prediction error and measurement error as potential confounders in detecting taste-based bias, but only to the extent that these confounders correlate

[8] Analogous to $\beta(z, r, v)$, we model $\lambda(z, r, v)$ additively, though our results could be easily adapted to a multiplicative formulation with minor modifications.

with defendant race. In Section 4, we show that prediction and measurement errors embedded in $\lambda(z, r, v)$ can invalidate the logic of the outcome test even if they are independent of defendant race.

2.3 Roy Model Representation

The decision model laid out above nests a family of Roy models, since heterogeneous potential outcomes enter directly into the decision maker's choice equation. At one extreme of the Roy spectrum, a decision maker in the original Roy (1951) model solely considers potential outcomes Y_d when choosing D, forging a very tight link between potential outcomes and decisions. In the pre-trial setting, such a model would require a judge to release only those defendants with the lowest causal effect of release on measured misconduct, with no other considerations. At the other end of the spectrum is the Generalized Roy Model (Björklund and Moffitt, 1987; Heckman and Vytlacil, 2005), in which potential outcomes are allowed to, but not required to, influence decisions, as factors other than the measured outcome Y_d can determine choices. In the pre-trial setting, this model would allow a judge to consider not only the effect of release on measured misconduct $Y_1 - Y_0$, but also defendant-level variation in detention costs C, bias β, and the phenomena captured in λ like prediction error ($\mathcal{E}_z \neq E$) and measurement error ($Y_d^* \neq Y_d$). A middle ground between these two extremes is the Extended Roy Model (Heckman and Vytlacil, 2007; D'Haultfœuille and Maurel, 2013), where other factors beyond potential outcomes are allowed to enter decisions but only in restricted ways, as we make precise below.

Generalized Roy Model

Without further restrictions, the judge's decision problem in (1) is a Generalized Roy Model (GRM), since it depends on expectations of not only the measured potential outcomes Y_d but also the (so far unrestricted) terms Y_d^*, C, and β. We can combine (2) and (3) and rearrange to make this representation explicit in our framework.

Definition 2.1 (Generalized Roy Model)**.** In the Generalized Roy Model (GRM), the decision rule of judge z facing a defendant characterized by (r, v) is given by

$$D_z = I\left\{E[\Delta|r, v] \leq \tau(z, r, v)\right\} \quad \text{where} \quad \tau(z, r, v) \equiv c(z, r, v) + \lambda(z, r, v) + \beta(z, r, v)\ . \tag{4}$$

In the GRM, the left side of the inequality is the mean effect of releasing a defendant with characteristics (r, v) on the measured misconduct outcome Y. The right side captures all other factors that influence the judge's decision of whether to release such a defendant. We collect these factors in the function $\tau(z, r, v)$, and refer to it as the perceived benefit of release. This benefit function $\tau(z, r, v)$ captures not only the expected detention costs $c(z, r, v)$ avoided by not detaining the defendant, but also any misalignment $\lambda(z, r, v)$ between the mean effect of release and the judge's prediction of it as defined in (3), as well as the taste for discrimination parameter $\beta(z, r, v)$.

The key elements of the GRM in (4) can therefore be summarized as follows. First, the cost of release $E[\Delta|r, v]$ is tied directly to defendant potential outcomes, which do not depend on the identity of the judge making the decision. Second, the benefit of release $\tau(z, r, v)$ contains all other factors that influence the judge's decision, including those that intrinsically depend on judge z via her beliefs and preferences. Third, both race R and non-race characteristics V enter both sides of the decision equation.

Remark 2.4. We could have started the description of the judge's decision problem with the GRM in (4) directly, without defining the optimization problem in (1). However, since many of our results in Sections 4 and 5 on the logical validity and econometric viability of the outcome test, as well as the task of defining bias in Section 3, hinge on properties of the benefit function $\tau(\cdot)$, the optimization problem is useful for showing explicitly how this function can arise as a composite of several conceptually distinct components, each of which may influence the properties of $\tau(\cdot)$. This derivation also allows us to demonstrate explicitly how our framework nests and relates to other papers in the outcome test literature.

Extended Roy Model

One of the key features of the GRM in Definition 2.1 is that both sides of the cost/benefit comparison—the expected misconduct cost on the left side, and the perceived benefits of release on the right side—are allowed to vary across all defendant characteristics in the judge's information set. This makes the GRM rich and flexible, but Sections 4 and 5 show that such flexibility introduces challenges to both the logical validity and econometric viability of outcome tests of bias.

We show these challenges can be nullified, however, by restricting the benefit function $\tau(\cdot)$ in

(4) to vary across only a subset of the defendant characteristics observed by the judge, narrowing the decision model from a Generalized Roy Model to an Extended Roy Model (ERM). Specifically, the expected misconduct cost $E[\Delta|r, v]$ on the left side of the decision inequality can continue to depend arbitrarily on all judge observables. However, the perceived benefit on the right side now depends only on the identity of the judge Z and the group stratification variable involved in the outcome test, namely race R, thus excluding all non-race defendant characteristics V from $\tau(\cdot)$.[9] Through the lens of the decision model, this means the expected cost of detention $c(z, r, v)$, the misalignment $\lambda(z, r, v)$ between the mean effect of release and the judge's prediction of it, and the taste for discrimination parameter $\beta(z, r, v)$ are all invariant across all non-race defendant characteristics v:

$$c(z, r, v) = c(z, r), \qquad \lambda(z, r, v) = \lambda(z, r), \quad \text{and} \quad \beta(z, r, v) = \beta(z, r) \ . \tag{5}$$

This restriction leads to the Extended Roy Model, as defined below.

Definition 2.2 (Extended Roy Model)**.** The decision rule of judge z in the Extended Roy Model (ERM) is given by

$$D_z = I\left\{E[\Delta|r, v] \leq \tau(z, r)\right\} \quad \text{where} \quad \tau(z, r) \equiv c(z, r) + \lambda(z, r) + \beta(z, r) \ . \tag{6}$$

Remark 2.5. The restrictions in (5) are sufficient for $\tau(z, r, v)$ not to depend on v. Of course, there exist functional forms of $c(z, r, v)$, $\lambda(z, r, v)$, and $\beta(z, r, v)$ such that $\tau(z, r, v)$ may not depend on v even when $c(z, r, v)$, $\lambda(z, r, v)$, and $\beta(z, r, v)$ do, but we do not view such specific cases as particularly insightful and so do not devote special attention to them.

Remark 2.6. In this framework, we could write the original Roy (1951) model as $D = I\{\Delta \leq 0\}$, augmented to include a fixed threshold as $D = I\{\Delta \leq \tau\}$. The Extended Roy Model thus extends the original Roy model in two ways: it allows the decision maker to have ex ante uncertainty over the potential outcomes Δ, and it allows the threshold τ to vary across different decision makers and

[9] More generally, the Extended Roy Model allows all variables commonly observed by the judge and the econometrician to drive variation in the benefit function $\tau(\cdot)$, but eliminates any variation in $\tau(\cdot)$ from variables observed by the judge but not the econometrician. We discuss in Section 5.1 how this accommodates empirical settings with observed covariates X and outcome tests of bias against other characteristics besides race.

defendant races (as well as any covariates X observed by both the judge and the econometrician, which we suppress here but discuss in Section 5.1).

Intuitive Comparison of the Models

The distinguishing feature of the ERM in (6) relative to the GRM in (4) is the exclusion of non-race defendant characteristics V from the benefit function $\tau(\cdot)$. This distinction has important implications for decision maker behavior. In the ERM, fixing z and r fixes the value of the benefit function $\tau(z, r)$. Thus, a given ERM judge z facing a pool of defendants of the same race r acts as if she sets a fixed cutoff of allowable misconduct risk equal to this fixed benefit value, and then releases all defendants with expected misconduct effects below this cutoff and detains all defendants above it. This cutoff can vary across different judges, and across different defendant races for a given judge, but is fixed for all defendants of a given race facing a given judge. Judge z's variation in release decisions within this pool of defendants is thus entirely driven by variation in expected measured misconduct, which in turn is driven by variation in non-race defendant characteristics V. It is thus important to observe that non-race defendant characteristics in the ERM play a starkly asymmetric role in a judge's cost-benefit analysis: such characteristics can shift the expected cost of release in terms of measured misconduct, but cannot shift any of the other decision factors collected in the benefit function, including detention costs, bias, prediction errors, and measurement errors.

The GRM, by contrast, does not impose this asymmetry: non-race defendant characteristics are allowed to shift both costs and benefits. Specifically, allowing the benefit function $\tau(z, r, v)$ to vary with V allows the judge to perceive different benefits of releasing defendants who share the same race but differ in their non-race characteristics, like employment status and number of dependents. Unlike in the ERM, the GRM would therefore allow a judge to set a higher cutoff of permissible misconduct risk among employed black defendants with children, for example, compared to unemployed black defendants without children. We next show how these modeling choices interact with the task of defining decision maker bias.

Remark 2.7. As we discuss in detail in Section 4.5, the family of Roy models we have laid out in this section underlie nearly all of the existing literature on outcome tests of bias. Arnold et al. (2018), for example, write down a model of bail judge behavior equivalent to the GRM in (4), while

Anwar and Fang (2006)'s model of police officer behavior in traffic stops is a special case of the ERM in (6). Knowles et al. (2001) consider a two-sided model of police behavior with endogenous driver responses, but the decision model of police officers—whose behavior is the object of interest in the outcome test—again represents a special case of the ERM.

3 Defining Bias

Section 2 laid out a family of decision models that describe the behavior of decision makers. In this section, we use these models to define racial bias and its absence, and discuss how our definitions relate to and nest notions of bias from the previous literature.

3.1 Absence of Racial Bias

The parameter $\beta(z, r, v)$ in the decision model (2) captures judge z's personal taste for releasing a defendant of race r and non-race characteristics v, distinct from considerations of misconduct risk and detention costs. It is natural, then, to first define the absence of racial bias in terms of this parameter as follows.

Definition 3.1 (Absence of racial β-bias)**.** We say judge z is racially β-unbiased if

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

This definition implies that a judge is racially unbiased if she has equal preferences for white and black defendants who share the same non-race characteristics v. Through its dependence on z and v, Definition 3.1 nests notions of unbiasedness from previous papers in the literature. The bias parameter in Knowles et al. (2001), for example, does not vary across decision makers (z here) or any non-race characteristics (v here) of drivers (defendants), which in turn simplifies their definition of unbiasedness. In Anwar and Fang (2006, Definition 1), bias is allowed to vary across decision maker race, but again not across drivers' (defendants') non-race characteristics. Definition 3.1 is similar to the definition of unbiasedness proposed by Brock et al. (2012, Equation (13)), where preferences are required to be invariant by race for each value of the non-race characteristics observed by the decision maker.

Defining absence of bias in terms of the preference parameter $\beta(z, r, v)$ is not the only option in a Generalized Roy decision model like (4). For example, Arnold et al. (2018, Definition 1) define absence of bias via the broader benefit function $\tau(\cdot)$ in (4), instead of the specific taste for discrimination parameter $\beta(\cdot)$, which is just one component of $\tau(\cdot)$. To contrast with Definition 3.1, we refer to absence of bias though the function $\tau(z, r, v)$ as "absence of racial τ-bias," and formally define it as follows.

Definition 3.2 (Absence of racial τ-bias)**.** We say judge z is racially τ-unbiased if

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

Definitions 3.1 and 3.2 are conceptually and mathematically different unless further restrictions are imposed. In order to appreciate the differences, it is helpful to recall that the benefit function $\tau(\cdot)$ captures three different terms:

$$\tau(z, r, v) \equiv c(z, r, v) + \lambda(z, r, v) + \beta(z, r, v) \ . \tag{7}$$

Two alternative interpretations immediately arise. First, one may interpret the two notions of unbiasedness in Definitions 3.1 and 3.2 as equivalent under the implicit assumption that the other two confounding functions, $c(z, r, v)$ and $\lambda(z, r, v)$, are absent, or at least independent of race. If that is the case, then any dependence of $\tau(\cdot)$ on race must come from the taste for discrimination parameter $\beta(\cdot)$, and the two notions would coincide. Second, one may proceed by being agnostic about the distinct mechanisms comprising the benefit function $\tau(\cdot)$ and define bias with a broad brush that goes beyond the taste for discrimination parameter $\beta(\cdot)$. Under this broader interpretation, for a judge to be labeled unbiased would require not only the personal taste for discrimination (β) to be race invariant; it would also require prediction/measurement errors (λ) to be independent of race, and require the expected cost of detention (c) to be independent of race.[10] Since these are a broader set of requirements, if judge z is racially τ-unbiased in the sense of Definition 3.2, she would also be racially β-unbiased in the sense of Definition 3.1, but the converse does not hold.

Definitions 3.1 and 3.2 simplify when preferences $\beta(z, r, v)$ and total release benefits $\tau(z, r, v)$,

[10] Analogous to Remark 2.5, throughout the paper we abstract away from functional singularities that may make the function $\tau(\cdot)$ race invariant even when its three components are not.

respectively, do not depend on non-race characteristics v. In that case, judge z is racially unbiased in the sense of Definition 3.1 when

$$\beta(z, r) = \beta(z) \ ,$$

and is racially unbiased in the sense of Definition 3.2 when

$$\tau(z, r) = \tau(z) \ .$$

With few exceptions, notably Brock et al. (2012) and Arnold et al. (2018), the literature on outcome tests typically proceeds under such restrictions, assuming (explicitly or implicitly) that decision maker bias is solely a function of defendant race, and thus invariant to all other non-race characteristics. Persico (2009, Equation (3)), for example, explicitly imposes such a restriction at the outset under the premise that it simplifies identification. Other papers, including Knowles et al. (2001) and Anwar and Fang (2006), do not highlight this restriction but impose it in their frameworks, which in turn is one of the points raised by Brock et al. (2012).

Recall that in order for $\tau(z, r, v)$ not to depend on v, one needs to restrict the Generalized Roy Model to an Extended Roy Model. Defining racially unbiased behavior in the ERM is therefore simpler than in the GRM, but requires a more restrictive model of how decision makers treat individuals with different *non-race* characteristics.

3.2 Types of Racial Bias

The subtleties associated with the preference parameter $\beta(z, r, v)$ generally depending on non-race characteristics and generally differing from the benefit function $\tau(z, r, v)$ are also present when formally defining the presence of bias. We start with the following set of definitions.

Definition 3.3 (Racial β-bias)**.**
We say judge z is *locally* β-biased against black defendants if

$$\beta(z, w, v) \geq \beta(z, b, v) \text{ for all } v \in \mathcal{V} \text{ with strict inequality for some open subset } \mathcal{V}_{\text{si}} \subseteq \mathcal{V} \ . \qquad (8)$$

We say judge z is *globally* β-biased against black defendants if

$$\beta(z, w, v) > \beta(z, b, v) \text{ for all } v \in \mathcal{V} \ . \tag{9}$$

Definition 3.3 labels judge z as biased against black defendants if she prefers white defendants over black defendants who have the same non-race characteristics v. As with Definition 3.1, the presence of non-race characteristics v introduces layers that are not present when bias only depends on (z, r), as in Knowles et al. (2001), Anwar and Fang (2006), and Persico (2009), among others. Our definition distinguishes between two degrees of bias, local and global. Decision makers may also exhibit "crossing" in their racial preference, i.e. $\beta(z, w, v) > \beta(z, b, v)$ for v in some subset of $\mathcal{V}$ and $\beta(z, w, v) < \beta(z, b, v)$ for v in some other subset of $\mathcal{V}$. In our formal results below, we refer to such a decision maker as "unclassified," which we define as any behavior that cannot be labeled as unbiased, locally biased, or globally biased.

As with the absence of bias, the presence of bias could alternatively be defined in terms of the overall benefit function $\tau(z, r, v)$ instead of its subcomponent $\beta(z, r, v)$.

Definition 3.4 (Racial τ-bias)**.** We say judge z is locally τ-biased or globally τ-biased against black defendants if (8) or (9), respectively, hold with $\tau(z, r, v)$ replacing $\beta(z, r, v)$.

Such a definition again admits two interpretations. One may interpret the two notions of bias in Definitions 3.3 and 3.4 as equivalent under the implicit assumption that the other two confounding components of $\tau(z, r, v)$—$c(z, r, v)$ and $\lambda(z, r, v)$—are independent of race. Alternatively, one may interpret bias in terms of $\tau(z, r, v)$ in a broad sense that goes beyond the taste for discrimination parameter $\beta(\cdot)$. Under this broader interpretation, there are multiple ways for a judge to be labeled as τ-biased: she may directly harbor a personal taste for discrimination (captured by β), but even if not, she may make prediction/measurement errors (λ) that vary by race, or expect the costs of detention (c) to vary by race, all of which would fall under the broader umbrella of τ-bias. A prominent example of τ-bias in the literature is Arnold et al. (2018, Definition 1), which defines judge j as racially biased against black defendants if $t_w^j(v_i) > t_b^j(v_i)$, where the function $t_r^j(v)$ is $\tau(z, r, v)$ in our notation, with judge j corresponding to our z and i indexing defendants. Since non-race defendant characteristics v_i are identical on both sides of the inequality (confirmed in the text of the definition), Definition 1 as published in Arnold et al. (2018) corresponds to our definition

of globally τ-biased in Definition 3.4.[11]

Definitions 3.3 and 3.4 simplify considerably when $\beta(z, r, v)$ and $\tau(z, r, v)$ do not depend on v. In that case, judge z is β-biased against black defendants if

$$\beta(z, w) > \beta(z, b) \ , \tag{10}$$

and is τ-biased against black defendants if

$$\tau(z, w) > \tau(z, b) \ . \tag{11}$$

It follows immediately that in the Extended Roy Model, there is no distinction between local and global bias. Furthermore, the possibility of "crossing" previously discussed does not arise and, as a result, judges cannot be labeled as unclassified. This again highlights a tradeoff between simpler definitions of bias (ERM) and richer models of decision making (GRM), which is important to keep in mind when comparing definitions of bias across the literature, studying the logical validity and econometric viability of outcome tests, and formulating new tests of bias in decision making.

4 Logical Validity of Outcome Tests of Bias

We now describe the outcome test, which aims to detect decision maker bias by comparing the post-decision outcomes of marginal individuals from different groups. In this section, we abstract away from the challenges of identifying marginal outcomes from data, and ask a more fundamental question: under what conditions are marginal outcome differences, even if known perfectly, actually informative about decision maker bias? We present two sets of results on the logical validity of the outcome test. The first set of results show that the test is logically invalid in the context of the Generalized Roy Model. We show this by generating broad classes of counterexamples which prove that the outcome test may conclude a judge is unbiased even if she is racially biased, or conclude bias even if the judge is racially unbiased, or even conclude bias against one race when the judge is biased against another race. The second set of results establish logical validity of the outcome

[11] In an unpublished correction appendix (Arnold et al., 2020), the authors change their definition of bias. We discuss this new definition in Section 4.5.

test when restricting decision maker behavior to the Extended Roy Model, as this restriction shuts down the confounding channels present in the Generalized Roy Model.

4.1 The Outcome Test of Racial Bias

To properly define the outcome test, we start with a precise notion of marginal defendants. For a given judge z and defendant race r, we denote by $v^*_{z,r} \in \text{int}(\mathcal{V})$ a value of v such that

$$E[\Delta | r, v^*_{z,r}] = \tau(z, r, v^*_{z,r}) \ . \tag{12}$$

That is, defendants of race r with non-race characteristics equal to $v^*_{z,r}$ are marginal for judge z in the sense that the judge acts as if the costs and benefits of releasing such a defendant exactly offset each other. An implicit minimal assumption we keep throughout our analysis is that such marginal values of v exist. For now, we also assume $v^*_{z,r}$ is unique for each (z, r) for simplicity, and discuss further complications from multiple crossing points in Section 5.

With this notation in hand, the marginal outcome test proceeds as follows. The test infers that judge z is racially biased against black defendants if

$$E[\Delta | w, v^*_{z,w}] > E[\Delta | b, v^*_{z,b}] \ , \tag{13}$$

i.e. if judge z's marginal white defendants have a higher mean effect of release on misconduct than her marginal black defendants. In this first case, the test is based on the premise that racial differences in misconduct outcomes among a judge's marginal defendants indicate that the judge is racially biased, specifically against the race with the lower marginal misconduct effect, as a result of holding that group to a stricter standard. Likewise, the test concludes there is no evidence of racial bias if

$$E[\Delta | w, v^*_{z,w}] = E[\Delta | b, v^*_{z,b}] \ , \tag{14}$$

i.e. if judge z's marginal white and marginal black defendants have equal mean effects of release on misconduct. In this second case, the test is based on the premise that equal misconduct outcomes among a judge's marginal white and marginal black defendants indicates that the judge is unbiased, holding each race to the same standard.

4.2 Logical Invalidity in the GRM

Our first result below shows that the marginal outcome test is logically invalid in the context of the GRM. At a conceptual level, the difference in the misconduct outcomes of marginally released black defendants versus marginally released white defendants is uninformative about whether the judge's cutoff depends directly on race because the GRM allows this cutoff to also vary with non-race defendant characteristics. We formally present this result in layers, by considering an increasing number of restrictions on the functions $E[\Delta|r,\cdot]$ and $\tau(z,r,\cdot)$ that enter the decision in (4). For simplicity, we focus on the case where v is a scalar, though this is not necessary for our results as we discuss in Appendix A. The intuition of our results is captured in Figure 1, which we discuss in detail after presenting the theorem.

To state the theorem, let $\mathcal{F}$ denote the space of all pairs of functions $E[\Delta|r,\cdot]$ and $\tau(z,r,\cdot)$ mapping $\mathcal{V}$ to $\mathbf{R}$. Then, for $\mathcal{V}^0$ being a non-empty open subset of $\mathcal{V}$, let $\mathcal{F}^{\mathrm{m}}(\mathcal{V}^0)$ denote the space of functions $E[\Delta|r,\cdot]$ and $\tau(z,r,\cdot)$ mapping $\mathcal{V}$ to $\mathbf{R}$ that are monotone (increasing or decreasing) on $\mathcal{V}^0$, and let $\mathcal{F}^{\mathrm{cm}}(\mathcal{V}^0)$ denote the subset of $\mathcal{F}^{\mathrm{m}}(\mathcal{V}^0)$ where these functions are also continuous. With this notation, $\mathcal{F}^{\mathrm{m}}(\mathcal{V})$ is the space of monotone functions, and $\mathcal{F}^{\mathrm{cm}}(\mathcal{V})$ is the space of continuously monotone functions. Definitions A.1-A.3 provide formal definitions of each of these spaces, including a few minor regularity conditions.

The space of functions $\mathcal{F}^{\mathrm{m}}(\mathcal{V}^0)$ is quite general and flexible enough to accommodate a variety of cases, including step functions. The space does not restrict the functions outside $\mathcal{V}^0$, but working with $\mathcal{F}^{\mathrm{m}}(\mathcal{V})$ the restrictions become global. The space of functions $\mathcal{F}^{\mathrm{cm}}(\mathcal{V})$ is the most restrictive one, requiring both $E[\Delta|r,\cdot]$ and $\tau(z,r,\cdot)$ to be continuously monotone functions of v on the entire domain $\mathcal{V}$.

Our first result can then be stated as follows.

Theorem 4.1. Consider the GRM in Definition 2.1, a given judge $z \in \mathcal{Z}$, and let $v^*_{z,r}$ be defined as in (12). Then, for each of the following four cases,

(i) Judge z is racially τ-unbiased,

(ii) Judge z is globally τ-biased against black or white defendants,

(iii) Judge z is locally τ-biased against black or white defendants,

(iv) Judge z is unclassified,

there exist functions $E[\Delta|r,\cdot]$ and $\tau(z,r,\cdot)$ in $\mathcal{F}$ such that the difference

$$E[\Delta|w, v^*_{z,w}] - E[\Delta|b, v^*_{z,b}] \tag{15}$$

could be positive, negative, or zero. The result continues to hold when $\mathcal{F}$ is replaced by $\mathcal{F}^{\mathrm{m}}(\mathcal{V}^0)$, $\mathcal{F}^{\mathrm{cm}}(\mathcal{V}^0)$, or $\mathcal{F}^{\mathrm{cm}}(\mathcal{V})$.

The theorem has multiple conclusions. First, the marginal outcome test may conclude bias even if the judge is racially unbiased. Second, the outcome test may conclude no bias even if the judge is locally or globally racially biased. Third, the outcome test may conclude bias in the opposite direction; for example, it may conclude bias against white defendants when the judge is locally or globally racially biased against black defendants. Importantly, the second and third conclusions continue to hold in the locally racially biased case regardless of whether $v^*_{z,r} \in \mathcal{V}_{\mathrm{si}}$ or not, where $\mathcal{V}_{\mathrm{si}}$ is the subset in Definition 3.4 in which the judge strictly prefers one race over the other. Finally, the outcome test may conclude bias in any direction, or conclude no bias, in the case where the judge is unclassified, where by unclassified we mean that the function $\tau(z,w,v)$ may be higher or lower than $\tau(z,b,v)$ depending on the value of v (i.e., they may cross at possibly multiple points). Each of these results hold regardless of whether the functions $E[\Delta|r,\cdot]$ and $\tau(z,r,\cdot)$ are left mostly unrestricted (as when they belong to the space $\mathcal{F}$), or are assumed to be well-behaved in some local area $\mathcal{V}^0$, or are assumed to be well-behaved globally (as when they belong to $\mathcal{F}^{\mathrm{cm}}(\mathcal{V})$), which demonstrates these results do not rely on pathological cases of cost and benefit functions. We relegate the proof of Theorem 4.1 to Appendix A and offer instead a graphical interpretation of its content in the next subsection.

Remark 4.1. Given the level of generality of Theorem 4.1, it is not difficult to see that a similar result applies when τ-bias is replaced with β-bias, as defined in Definitions 3.1 and 3.3. That is, differences in the misconduct outcomes of marginally released black defendants versus marginally released white defendants are uninformative about whether a GRM judge is β-biased or not. The reason is that any restriction imposed on $\beta(\cdot)$ will necessarily lead to $\tau(\cdot)$ falling into one of the four cases considered in Theorem 4.1, and from there the result will follow. Note that this includes

restricting $\beta(\cdot)$ to be invariant to non-race characteristics, thus allowing for the simpler definition of (racial-only) taste-based bias used by much of the previous literature, but leaving the other components of $\tau(\cdot)$ unrestricted. Thus, the results in Theorem 4.1 are not sensitive to any of these alternative definitions of bias.

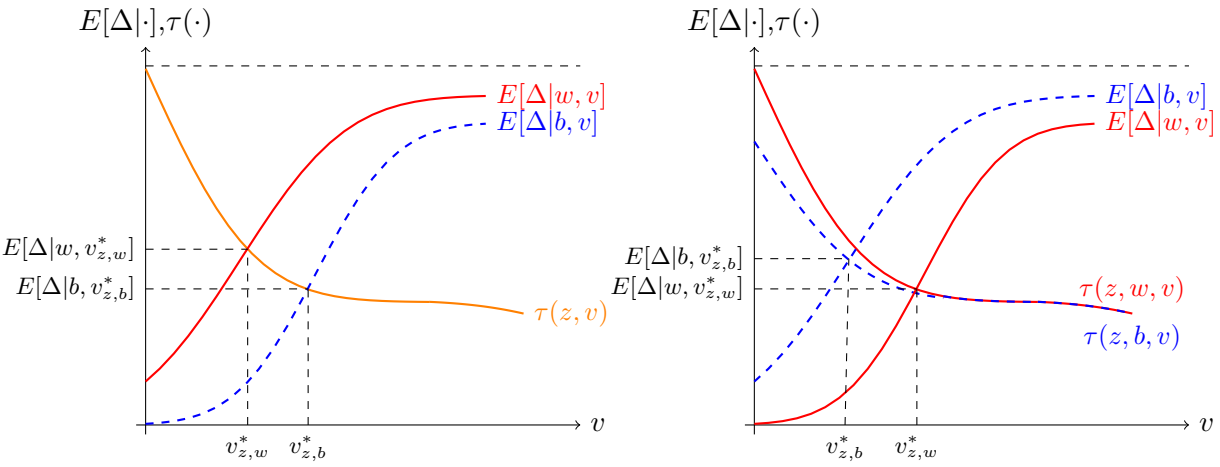
Remark 4.2. In the context of police searches, Brock et al. (2012) show that the arguments underlying the comparison of *average* post-decision outcomes across groups in both Knowles et al. (2001) and Anwar and Fang (2006) break down if the search cost function (our τ) depends on non-race characteristics that are unobserved by the analyst (our v). This observation is similar in spirit to our Theorem 4.1, but Brock et al. (2012), reacting to the state of the literature at the time, study the mapping from *average* outcome comparisons to racial bias, and thus focus on the inframarginality complications that arise when the distribution of V differs by race, since average comparisons integrate over these distributions. In contrast, our results in Theorem 4.1 concern the more fundamental mapping from *marginal* outcomes to bias, which only involves individuals with specific values of v—namely those with $v = v^*_{z,r}$, i.e. marginal individuals of race r facing decision maker z. None of our results, therefore, are affected by whether V is distributed differently across races, which is a central concern in Brock et al. (2012) and the entire literature on average-based outcome tests.

Remark 4.3. In a context where the analyst is interested in testing the null hypothesis H_0 : "judge z is racially τ-unbiased," an immediate implication of Theorem 4.1 is that the marginal outcome test will not control size and will have no power against certain alternatives. This point holds independently of how well the analyst is able to estimate the difference $E[\Delta|w, v^*_{z,w}] - E[\Delta|b, v^*_{z,b}]$.

4.3 Graphical Representation of the Formal Results

In Figure 1, we explain the intuition behind the formal results of logical invalidity of the outcome test in the GRM by drawing examples of cost and benefit functions for black and white defendants facing a given judge. The functions $E[\Delta|r, \cdot]$ and $\tau(z, r, \cdot)$ in the figures belong to $\mathcal{F}^{\text{cm}}(\mathcal{V})$ to illustrate how logical invalidity obtains even with globally well-behaved cost and benefit functions.

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 τ-unbiased according to Definition 3.2. However, the cost function



(a) Intuition with no racial bias

(b) Intuition with local racial bias

Figure 1: Intuition behind Theorem 4.1

$E[\Delta|r,v]$ does vary with race, consistent with the GRM in (4). In this case, $E[\Delta|w,v]$ is greater than $E[\Delta|b,v]$ for all v. So, even though judge z is not biased against black defendants, the fact that black defendants have a lower expected effect of release on misconduct for a given value of the non-race characteristics v leads to the marginal black defendant having a strictly lower misconduct effect than the marginal white defendant, $E[\Delta|b,v^*_{z,b}] < E[\Delta|w,v^*_{z,w}]$. Thus, Figure 1a illustrates a case in which condition (i) of Theorem 4.1 holds with (15) being positive. That is, the marginal outcome test may erroneously conclude racial bias when in fact the judge is racially τ-unbiased.

Figure 1b illustrates a case where race enters both the benefit function $\tau(z,r,v)$ and the cost function $E[\Delta|r,v]$. In this case, $E[\Delta|b,v]$ is greater than $E[\Delta|w,v]$ for all v, while $\tau(z,w,v)$ is strictly greater than $\tau(z,b,v)$ for low values of v, consistent with condition (iii) in Theorem 4.1. Even though judge z is locally biased against black defendants in the sense of Definition 3.4, the fact that black defendants have a sufficiently higher expected effect of release on misconduct for a given value of the non-race characteristics v leads to the marginal black defendant having a strictly higher misconduct effect than the marginal white defendant, $E[\Delta|b,v^*_{z,b}] > E[\Delta|w,v^*_{z,w}]$. Thus, Figure 1b illustrates a case in which condition (iii) of Theorem 4.1 holds with (15) being negative. In other words, the marginal outcome test may erroneously conclude bias against one race when in fact the judge is racially biased against another race.

4.4 Logical Validity in the ERM

The logical invalidity of the marginal outcome test arises because the GRM defined in Definition 2.1 does not require an unbiased judge to equalize misconduct outcomes across marginal white and marginal black defendants. This requirement, however, turns out to hold in the context of the ERM defined in Definition 2.2. Our next result shows that the difference in the misconduct outcomes of marginally released black defendants versus marginally released white defendants immediately informs whether an ERM judge's cutoff depends directly on race, since a given judge's cutoff *only* depends on race in the ERM.[12] In other words, the marginal outcome test is indeed logically valid when the decision model conforms to the ERM, as formalized in the next result.

Theorem 4.2. Consider the ERM in Definition 2.2 for a given judge z. Then, the difference in the misconduct outcomes of marginally released black defendants versus marginally released white defendants equals the difference in the expected benefit function $\tau(z, r)$ between black and white defendants, i.e.,

$$E[\Delta|w, v^*_{z,w}] - E[\Delta|b, v^*_{z,b}] = \tau(z, w) - \tau(z, b) \ . \tag{16}$$

The outcome test is then logically valid in the sense that

$$\begin{aligned} E[\Delta|w, v^*_{z,w}] > E[\Delta|b, v^*_{z,b}] \iff \tau(z, w) > \tau(z, b) \quad \text{and} \\ E[\Delta|w, v^*_{z,w}] = E[\Delta|b, v^*_{z,b}] \iff \tau(z, w) = \tau(z, b) = \tau(z) \ . \end{aligned}$$

The logical validity of the outcome test is fully restored in the ERM: it would correctly conclude (no) bias if judges are in fact (un)biased. The intuition for this follows from Figure 1 by letting the functions $\tau(z, r, v)$ be flat, i.e. invariant across non-race characteristics v, in which case each must intersect the two cost functions at the same vertical value. In fact, Theorem 4.2 not only shows that outcome tests are logically valid in the ERM, but also shows through (16) that the magnitude of judge z's racial τ-bias, $\tau(z, w) - \tau(z, b)$, is identified by the magnitude of the marginal outcome difference $E[\Delta|w, v^*_{z,w}] - E[\Delta|b, v^*_{z,b}]$. In the language of Becker (1957), this difference succeeds in revealing judge z's "willingness to pay" to discriminate, measured in terms of excess misconduct: the amount of additional misconduct among white defendants that the judge is willing to allow in

[12] Potentially conditional on covariates X observed by both the judge and the analyst, as we discuss in Section 5.1.

order to express her τ-bias against black defendants, or vice versa.[13]

Remark 4.4. While the ERM can deliver a logically valid test of bias in terms of $\tau(\cdot)$, consistent with our τ-bias Definitions 3.2 and 3.4, the decomposition

$$\tau(z,r) = \beta(z,r) + \lambda(z,r) + c(z,r)$$

is still present in the ERM. Without further restrictions, then, the three components entering $\tau(z,r)$—preferences, prediction/measurement errors, and detention costs (i.e. other judge considerations beyond misconduct)—are indistinguishable from each other in the marginal outcome difference $E[\Delta|w, v^*_{z,w}] - E[\Delta|b, v^*_{z,b}]$. This illustrates how difficult it is to isolate the taste component $\beta(z,r)$ from other confounding factors captured by $\lambda(z,r)$ and $c(z,r)$ without direct measures of each (e.g., Bohren et al., 2021) or further theoretical restrictions like $\lambda(z,r) = \lambda(z)$ and $c(z,r) = c(z)$.

Remark 4.5. The ERM allows marginal black and white defendants to have different values of their non-race characteristics, i.e., $v^*_{z,w} \neq v^*_{z,b}$, even when the judge is racially τ-unbiased. In Figure 1(a), for example, if judge z satisfied both the ERM and τ-unbiasedness such that the benefit function were a constant $\tau(z)$, we would still have $v^*_{z,w} < v^*_{z,b}$. This would be consistent with statistical discrimination: the judge would release a black defendant with non-race characteristics v somewhere between $v^*_{z,w}$ and $v^*_{z,b}$, since $E[\Delta|b,v] < \tau(z)$ for $v \in (v^*_{z,w}, v^*_{z,b})$, but would detain an otherwise identical white defendant, since $E[\Delta|w,v] > \tau(z)$ for the same value of v. This would be entirely driven by racial differences in the outcome function $E[\Delta|r,v]$ rather than bias, as $\tau(z)$ is constant across race (and all other defendant characteristics as well).

4.5 Discussion and Implications for the Literature

The main takeaway from Theorem 4.1 and Remark 4.1 is that sufficiently flexible models of decision making do not deliver logically valid outcome tests of bias under any of the definitions of bias discussed in Section 3. Theorem 4.2 offers a clear solution, in the form of restricting the model of

[13]Since τ-bias $\tau(z,w) - \tau(z,b)$ is specific to a given decision maker z, these results readily apply to outcome tests of in-group bias, which aim to examine whether the presence and magnitude of bias differs across different decision makers grouped by some observable characteristic like race or gender (e.g., Shayo and Zussman, 2011; Anwar et al., 2012; Knepper, 2018; Ash et al., 2022).

decision maker behavior to an Extended Roy Model. Together, these results provide a clarifying lens through which to understand the existing literature on outcome tests, and a guide for empirical research that aims to test for or quantify bias in decision making.

Most immediately, our results tell a cautionary tale for approaches that simultaneously attempt to maintain the flexibility of the Generalized Roy Model, which is a workhorse in modern applied economics (Heckman, 2010), while also attempting to test for bias. An illustrative case study is Arnold et al. (2018), who write down a model of bail judge decision making that satisfies our Definition 2.1 of a GRM. By studying a decision environment that features exogenously assigned decision makers who vary progressively in their decision tendencies, Arnold et al. (2018) build on a rapidly growing literature that employs these institutional features to identify causal consequences of an array of policy-relevant decisions.[14] An important innovation of Arnold et al. (2018) is their recognition that these institutional features may also help address two perennial identification challenges faced by the discrimination literature: selection bias (as judges are randomly assigned to cases) and inframarginality (as judges vary progressively in their decision tendencies, potentially facilitating marginal comparisons). 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). Despite these methodologically amenable institutional features, the logic of the outcome test proposed in the GRM framework of Arnold et al. (2018) is invalidated by Theorem 4.1, as the definition of racial bias in Arnold et al. (2018) corresponds to τ-bias in our Definition 3.4, as noted in Section 3.2.

An important lesson to draw from this case is that addressing selection and inframarginality is not sufficient for a valid outcome test. The reason is that selection and inframarginality are issues of identification, which we turn to in Section 5. These issues arise when attempting to use observed variation across decision makers and across cases within decision makers to identify the outcomes of a given decision maker's marginal cases. In contrast, the logical validity of the outcome test, which we have studied in this section, concerns the mapping between these marginal outcomes, even if known perfectly, and the bias of the decision maker. Theorem 4.1 reveals a lack of such a mapping for Generalized Roy decision makers under any of the definitions of bias in Section

[14]This includes 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).

3. Thus, regardless of whether a given decision environment features exogenous assignment of decision makers and progressive variation in decision tendencies, or whether a researcher develops other methods of dealing with selection and inframarginality, avoiding the implications of Theorem 4.1 must involve choosing a different definition of bias beyond those we discuss in Section 3, or choosing a different model of decision making that is less flexible than the GRM. We now discuss each of these possibilities in turn.

Choice of Bias Definition

In an unpublished correction appendix (Arnold et al., 2020), the authors of Arnold et al. (2018) respond to our result in Theorem 4.1. To restore the logical validity of their proposed outcome test, Arnold et al. (2020) maintain their Generalized Roy decision model but change the definition of bias. Under the new bias definition, judge z is racially biased against black defendants if

$$\tau(z, w, v^*_{z,w}) > \tau(z, b, v^*_{z,b}), \tag{17}$$

i.e. if judge z perceives greater benefits of releasing her marginal white defendant compared to her marginal black defendant.[15] By the definition of marginal white and black defendants, (17) is equivalent to $E[\Delta|w, v^*_{z,w}] > E[\Delta|b, v^*_{z,b}]$. Defining bias in this way therefore restores the logical validity of the outcome test, since the sign (and exact magnitude) of the marginal outcome comparison $E[\Delta|w, v^*_{z,w}] - E[\Delta|b, v^*_{z,b}]$ is identical that of $\tau(z, w, v^*_{z,w}) - \tau(z, b, v^*_{z,b})$.

To understand the limitations of this new bias definition, it is important to recognize that (17) compares a judge's perceived benefits of releasing black and white defendants with *different* non-race characteristics, since $v^*_{z,w} \neq v^*_{z,b}$ in general. This contrasts with our definitions of bias in Generalized Roy Models in Section 3, which hold v fixed, as well as with the seminal discrimination framework of Becker (1957) that Arnold et al. (2018) cite as motivation for their outcome test. Becker (1957) analyzed implications of bias against minority workers who were otherwise *identical* to non-minority workers; if they were instead "imperfect substitutes, they may receive different wage rates even in the absence of discrimination" (p. 17). The definition in (17) is also silent about

[15] In a different setting, Appendix B of Dobbie et al. (2021) uses (17) to define taste-based bias among loan examiners. In the main text of Dobbie et al. (2021), however, the authors use a different, ERM-based definition of "incentive-based" bias that satisfies (11), our definition of τ-bias in the ERM.

how the functions $\tau(z, w, v)$ and $\tau(z, b, v)$ relate to each other for any other values of v beyond those that characterize judge z's marginal defendants. Thus, a judge labeled as biased under this revised definition could nonetheless satisfy $\tau(z, w, v) = \tau(z, b, v)$ for all $v \in \mathcal{V}$, i.e. be racially τ-unbiased according to Definition 3.2. We visualized such a case in Figure 1a: judge z's perceived benefit function excludes race entirely as an argument, yet the definition in (17) would label such a judge as biased. Conversely, a judge who is labeled as unbiased, or biased against white defendants, according to (17) could nonetheless satisfy $\tau(z, w, v) > \tau(z, b, v)$ for all $v \in \mathcal{V}$, i.e. satisfy the especially stark notion of global racial τ-bias against black defendants according to Definition 3.4.

Another recent example of avoiding Theorem 4.1 by studying target parameters other than the definitions of bias in Section 3 is Arnold et al. (2022). In their pre-trial release context, Arnold et al. (2022) define racial discrimination as "the average release rate disparity between white and Black defendants with identical misconduct potential." In our notation, and for $D = D_Z$, this involves comparing the release rates of white and black defendants facing a given judge $Z = z$ who have equal misconduct cost $\Delta = \delta$,

$$E[D|Z = z, R = w, \Delta = \delta] - E[D|Z = z, R = b, \Delta = \delta] \ , \tag{18}$$

and then averaging over the support of misconduct cost values δ. If this difference is positive, then the judge is more likely to release white defendants than black defendants with equal misconduct potential, and is classified as discriminating against black defendants.

Using an identification-at-infinity argument, Arnold et al. (2022) show how the average of (18) could be recovered without an explicit model of judge behavior. This leads the authors to describe their definition and test for discrimination as "model-free." The economic interpretation and policy implications of such a test, however, unavoidably depend on how judges are assumed to be making decisions and generating the data. If we are not willing to go beyond "model-free" quantities like (18), then we cannot distinguish among the wide range of mechanisms that determine these quantities. As we illustrate in Appendix D, this includes taste-based bias, prediction/measurement errors, and detention costs—together comprising $\tau(z, r, v)$ in our model—as well as statistical discrimination via differences in $E[\Delta|r, v]$ by race, and differences in the distribution of V conditional on $R = r$ and $\Delta = \delta$. Each of these mechanisms clearly has different implications for potential

policy interventions. In the latter part of their paper, Arnold et al. (2022) indeed end up imposing a model of judge decision making similar to the ERM in an attempt to distinguish among these distinct mechanisms.[16]

Choice of Decision Model

As an alternative to choosing target parameters that differ from the definitions of bias in Section 3, one may carefully specify a model of decision making that delivers a logically valid outcome test. Theorem 4.2 offers such a model, the Extended Roy Model. Interestingly, this model of decision making forms the basis of important prior work that aims to test for bias as defined in Section 3.

One set of studies, launched by Knowles et al. (2001), build on Becker (1957) and the large follow-on literature using differences in average outcomes across groups to test for discrimination, but pay careful attention to the threat to this literature posed by inframarginality bias: average outcomes need not be informative about conditions at the margin, due to group differences in unobservables among individuals away from the decision margin who are included in average comparisons. Knowles et al. (2001) address this issue by endogenizing the behavior of the individuals subject to the decision process—in this case, drivers deciding whether to carry contraband—and show how the equilibrium of such a model can discipline average outcomes to reveal marginal outcomes.[17]

While this kind of equilibrium model may seem distinct from our single decision maker framework in Section 2, we reiterate the point made earlier that inframarginality is an issue of identification, rather than logical validity. Importantly for the discussion at hand, the model of decision maker behavior in Knowles et al. (2001) compares the expected outcome, conditional on race and non-race characteristics observed by the decision maker, to a threshold that depends on race but *not* on non-race characteristics, exactly as in the ERM. Furthermore, bias ("prejudice") is defined as differences in these thresholds by race, as in our ERM-based definition of τ-bias in (11). Regard-

[16] As in the ERM, the model in Arnold et al. (2022)'s decomposition exercise features a threshold function that depends on the identity of the judge z and the race of the defendant r, but excludes non-race defendant characteristics v. On the other side of the decision model, judges form an expectation about defendant misconduct if released, which depends on defendant race and a noisy "signal" V drawn from a parameterized distribution that varies across judges and defendant races. Thus Arnold et al. (2022)'s decision model in our notation is $D_z = I\{\mathcal{E}_z[Y_1|r,v] \leq \tau(z,r)\}$, with parameterizations that yield a known functional form for $\mathcal{E}_z[Y_1|r,v]$.

[17] Other papers in this literature relying on equilibrating behavior include Persico (2002), Hernández-Murillo and Knowles (2004), and Persico and Todd (2006). See Persico (2009) for a review, as well as Dharmapala and Ross (2004), Dominitz and Knowles (2006), Manski (2006), and Bjerk (2007) for discussions and extensions.

less of the additional equilibrium structure imposed to identify marginal outcomes from average outcomes, then, the basic logic of the outcome test in Knowles et al. (2001) is secured by Theorem 4.2, in that the outcomes of marginal individuals are indeed informative about τ-bias when the decision model is an ERM.

A second set of studies, pioneered by Anwar and Fang (2006), relax the assumption of equilibrating driver responses to police decision rules by exploiting variation in decision rates and average outcomes across different decision makers.[18] Specifically, Anwar and Fang (2006) show that relative comparisons of the rates at which police officers of different races decide whether to search stopped drivers of each race, and the rates at which those searches find contraband, can deliver testable implications about racial differences in criminal behavior among drivers on the margin of being searched. Since Anwar and Fang (2006)'s model of police behavior satisfies our Definition 2.2 of an ERM (with different decision makers z distinguished by their race r_p), these marginal outcomes are in turn informative about decision maker bias by Theorem 4.2, as the definition of bias in Anwar and Fang (2006) also corresponds to our ERM-based definition of τ-bias in (11).

Remark 4.6. While the Extended Roy Model of decision maker behavior thus underpins much of the outcome test literature, the ERM is not strictly necessary for a logically valid outcome test of bias as defined in Section 3. Figure 1 suggests another potential solution: rather than restricting the benefit/threshold function $\tau(\cdot)$, the analyst could restrict the cost/outcome expectation function to be identical across race, i.e.

$$E[\Delta|r, v] = E[\Delta|v] \ . \tag{19}$$

In Appendix C, we show that this restriction can deliver a logically valid outcome test of global τ-bias, per Definition 3.4, if $E[\Delta|v]$ is strictly increasing in v and $\tau(z, r, v)$ is weakly decreasing in v. Intuitively, this restriction restores logical validity by assuming away any objective statistical relationship between race and potential outcomes, leaving τ-bias as the only channel in the decision model that could cause racial disparities in marginal outcomes. To justify such a restric-

[18] Another prominent example of this approach is Alesina and La Ferrara (2014), who study racial bias in death penalty sentencing by exploiting variation in reversal rates by higher courts across combinations of defendant race and victim race. Marx (2022) revisits Anwar and Fang (2006)'s dataset with an absolute, rather than relative, test of bias among police officers of different races.

tion, the analyst would need to not only rule out any direct causal effects of race on potential outcomes, but also require the decision maker's information set to encompass all determinants of potential outcomes that correlate with race, leaving race as an irrelevant predictor conditional on that information set.

Implications for Logically Valid Outcome Tests across Empirical Settings

We have highlighted two alternative pathways out of the negative result in Theorem 4.1. The first pathway—changing the parameter of interest to something beyond the definitions of bias in Section 3—amounts to asking a different research question. To the extent that alternative definitions like (17) and (18) capture relevant questions of interest, they can lead to logically valid outcome tests despite maintaining highly flexible models of decision maker behavior. On the other hand, if we are interested in testing for bias as defined in Section 3 and in prior work like Knowles et al. (2001) and Anwar and Fang (2006), then the other pathway out of Theorem 4.1 is required: restricting the model of decision making. Such restrictions need to be considered and defended on a case-by-case basis, depending on the empirical setting.

Consider first our running example of pre-trial release. Here, imposing the Extended Roy Model restores the logical validity of the outcome test (Theorem 4.2) by eliminating all variation in a given judge's perceived benefit of releasing defendants of the same race, leaving racial bias and expected misconduct as the sole sources of variation in a given judge's decisions. This restriction has important implications for the admissible behavior of bail judges, as well as for the measurement of the misconduct outcome. To justify it, the analyst would need to rule out any additional judge biases against any defendant characteristics unobserved by the analyst, like speech patterns, body weight, disabilities, and physical demeanor. Relatedly, the magnitude of a judge's racial bias must not vary across any such characteristics. Errors in judge predictions of misconduct, or in the analyst's measurement of misconduct, must also not correlate with any non-race defendant characteristics. Finally, the analyst must rule out any judge considerations beyond expected misconduct, like employment or family disruptions, that vary across defendants with different non-race characteristics. Violations of any of these restrictions would manifest as the benefit function $\tau(\cdot)$ varying with v, reopening the applicability of Theorem 4.1.

Studies of racial bias in pre-trial release like Arnold et al. (2018), Arnold et al. (2022), and Hull

(2021) often do consider complications to the outcome test posed by prediction error, measurement error, and omitted judge considerations, but only to the extent that these confounding factors correlate with defendant race. Such phenomena manifest as $\tau(\cdot)$ varying with r for reasons other than pure racial animus, represented in our framework as λ and c depending on r. But a novel and perhaps surprising lesson from our analysis is that the presence of any of these confounding factors can invalidate the outcome test even if they are completely independent of race—as Theorem 4.1 does not depend on the statistical relationship between R and V—as long as they manifest as $\tau(\cdot)$ varying with non-race characteristics. This is precisely why imposing an ERM delivers logical validity, as the ERM's exclusion of v from $\tau(\cdot)$ shuts down all such possibilities.

The ERM could permit logically valid outcome tests of bias across a range of settings beyond pre-trial release, with careful consideration paid to the credibility of the ERM's restrictions in each environment. We have already discussed how Knowles et al. (2001), Anwar and Fang (2006), and related papers employ the ERM to test for racial bias among police officers conducting traffic stops, spurring a lively literature discussing and debating the assumptions behind this approach.[19] Building on Becker (1993)'s suggestion of testing for bias in lending markets by studying borrower outcomes, Dobbie et al. (2021) derive an ERM from a principal-agent model of a loan officer who compares the short-run expected profitability of a loan application to a fixed threshold. The ERM may reasonably describe decision maker behavior in such a setting, as loan officers have a well-incentivized and well-measured objective and detailed data on prior applicants and outcomes.

In the seemingly distant realm of academic peer review, Card and DellaVigna (2020), Card et al. (2020), and Hengel and Moon (2020) study gender differences in citation outcomes through the lens of a model in which journal editors publish papers with expected citations above a threshold, which may vary by author gender and other analyst observables but not by factors unobserved by the analyst—again as in the ERM. The credibility of this model therefore hinges on the analyst observing and conditioning on all non-gender factors that may shift around the editor's threshold, putting aside the identification challenge of inframarginality bias from using average citations to try to infer the citation outcomes of marginally accepted papers.

In college admissions, an active literature on "mismatch" studies whether students from different

[19] E.g., Dharmapala and Ross (2004); Antonovics and Knight (2009); Dominitz and Knowles (2006); Bjerk (2007); Persico (2009); Brock et al. (2012).

demographic backgrounds reap different returns to attending more selective institutions.[20] Viewed as an outcome test, an analyst might consider whether outcome differences among marginally admitted applicants are informative about a college's preferences over applicant demographics (Bhattacharya and Shvets, 2022), given that differences in admission *rates* across groups may simply reflect selection on applicant characteristics unobserved by the analyst. Imposing an ERM in this setting to satisfy Theorem 4.2 would require ruling out any other college objectives—beyond the analyst's observed outcome, often degree completion or earnings—that vary with analyst-unobserved student characteristics.[21]

A final example is welfare programs with examiner discretion, like disability insurance, where an analyst might consider testing for bias among case examiners by comparing the subsequent labor market outcomes of marginally rejected applicants from different groups. Imposing the ERM in this setting would require an examiner to approve or reject applicants solely by comparing expected employment or earnings potential to a cutoff, where that cutoff may vary across the group membership potentially subject to bias but not across any other applicant or case characteristics excluded from the analyst's data set.

Regardless of the specific empirical setting, our results highlight that testing for bias according to the definitions in Section 3 requires the analyst to specify a model of decision maker behavior that forges a sufficiently tight link between potential outcomes and decisions. Otherwise, logical validity of outcome tests can fail. Writing down an explicit model of decision making is therefore a crucial part of testing for bias, as it clarifies exactly how the analyst is defining bias and what, if anything, the data can reveal about such bias. While this is not a new insight (e.g., Becker, 1957; Heckman, 1998; Knowles et al., 2001; Anwar and Fang, 2006), it is a useful reminder given that much of the empirical literature on discrimination does not employ explicit models of decision making, failing to clarify when and why observed disparities should be equated with decision maker bias.

[20] E.g., Arcidiacono and Lovenheim (2016); Bleemer (2022); Mountjoy and Hickman (2021).

[21] Legacy students, for example, may be more generous in their future donations to the college compared to non-legacies with the same earnings, incentivizing colleges to lower their admission threshold for applicants with legacy status (e.g., Arcidiacono et al., 2021).

5 Econometric Viability of Outcome Tests of Bias

The outcome test of bias requires the analyst to compare the outcomes of marginal individuals across groups, as formalized in Section 4.1. Our discussion up to this point focused on the logical validity of this test, which addressed the question of whether group differences in the outcomes of marginal individuals, even if perfectly known, are informative about the presence and magnitude of decision maker bias. In this section, we turn to the question of identifying these objects from data, as outcomes of marginal individuals are rarely observed directly by the analyst. In Section 5.1, we consider the best-case scenario of local instrumental variation across continuously distributed and exogenously assigned decision makers. In Section 5.2, we shift attention to other settings and data environments, including the empirically relevant situations of discrete instruments and no instruments.

5.1 Best-Case Scenario with Continuous and Exogenous Decision Makers

Depending on the setting, the support of decision makers Z could be modeled as a continuum in $\mathbf{R}$ or as a discrete set with few or many elements. In the best-case scenario of continuous and exogenous variation in Z, researchers typically invoke the marginal treatment effect (MTE) framework of Heckman and Vytlacil (2005) and use the local instrumental variables (LIV) estimand to try to recover the outcome test's parameters of interest as MTEs. In this subsection, we show that this approach delivers an econometrically viable outcome test under the ERM. In Appendix B, we show that the GRM, in contrast, not only fails to deliver a logically valid test, but also does not permit identification of the marginal outcomes of interest without additional restrictions. Importantly, imposing such restrictions to ensure identification of MTEs under the GRM does not address the issues of logical failure identified in Section 4. An analyst who does not impose the ERM may therefore successfully identify marginal outcome differences across groups that are nevertheless uninformative about decision maker bias as defined in Section 3.

Observed and Unobserved Variables

Let (Y, Z, D, R, X) be the random vector collecting all the variables observed by the analyst. As before, Z denotes a decision maker who is exogenously assigned to cases and makes a binary decision

$D = D_Z$ regarding an individual characterized by group membership R. The observed outcome of interest, Y, relates to the potential outcomes by the relationship $Y = DY_1 + (1 - D)Y_0$. For concreteness, we continue to ground our discussion in the setting of racial bias in pre-trial release, where Y is a measure of pre-trial misconduct and R denotes the defendant's race.

More generally, our analysis applies to any outcome variable $Y \in \mathbf{R}$, as well as any group stratification variable R taking finitely many values. With X denoting the set of covariates that are observed by both the judge and the analyst, the analyst could choose other elements of X as the group membership of interest for an outcome test, like gender, age, or nationality. In that case, R would represent this group membership of interest, and race would be an element of the covariate set X. All of the analysis then proceeds by conditioning on necessary elements of X (which we keep implicit in the notation) and conducting comparisons across values of R. Importantly, however, note that the analyst does not observe the random variable V that captures all of the non-race defendant characteristics that are only observed by the judge.[22]

Local IV Successfully Implements the Outcome Test under the ERM

We now show that under the ERM, the local IV (LIV) estimand identifies the marginal treatment effects of interest for the outcome test. The argument proceeds in two steps: (i) the ERM has a separable index model representation and thus ensures that LIV recovers an MTE for each race, and (ii) the ERM also ensures that these race-specific MTEs correspond to the parameters of interest in a logically valid outcome test.

To show (i), we map the decision model into the MTE framework by writing the general release decision in (4) as a latent index model,

$$D = I\Big\{\zeta(Z, R, V) \leq 0\Big\}\ , \tag{20}$$

where $\zeta(Z, R, V) \equiv E[\Delta|R, V] - \tau(Z, R, V)$ represents the net cost of release. Under the ERM, $\tau(\cdot)$

[22] While the dimension of V did not affect the results in Section 4, here we invoke results from the MTE framework that are better described for a scalar V. Most of the results could be extended to the vector case at the cost of more complex notation, but we prefer to focus on the case of a scalar V for the sake of clarity.

does not depend on V, yielding an index that is separable in Z and V:

$$\zeta(Z, R, V) \equiv E[\Delta|R, V] - \tau(Z, R) \ . \tag{21}$$

Separability of the latent index between the instrument and unobserved heterogeneity plays an important role in the identification of MTEs via LIV.[23] When the function $\zeta(Z, R, V)$ is separable in Z and V, as in the ERM, there is a convenient representation of the decision equation in terms of the propensity score $p(z, r) \equiv P\{D = 1|Z = z, R = r\}$ and a uniformly distributed random variable U_r. Specifically, let $F_r(\cdot)$ denote the CDF of $E[\Delta|R, V]$ conditional on $R = r$, which is assumed to be one-to-one. Then,

$$D = I\{\zeta(z, r, V) \leq 0\} = I\{F_r(E[\Delta|r, V]) \leq F_r(\tau(z, r))\} = I\{U_r \leq p(z, r)\} \ , \tag{22}$$

where $U_r \sim U[0, 1]$ and $p(z, r) = F_r(\tau(z, r))$.

With this representation, marginal defendants of race r facing judge z are characterized by $U_r = u^*_{z,r} \equiv p(z, r)$, and the marginal treatment effect of releasing such defendants is $E[\Delta|R = r, U_r = u^*_{z,r}]$.[24] Heckman and Vytlacil (2005) show that this MTE, defined with respect to U_r, is identified by the LIV estimand,

$$LIV(r, \tilde{p}) \equiv \frac{\partial}{\partial \tilde{p}} E[Y|R = r, p(z, r) = \tilde{p}] = E[\Delta|R = r, U_r = u^*_{z,r}] \ , \tag{23}$$

which establishes (i) above.

To show (ii), we return to the decision model in terms of V. Denote by $\{v^*_{z,r,j} : j \in J_{z,r}\}$ the set of marginal values of V for judge z and defendant race r such that $\zeta(z, r, v^*_{z,r,j}) = 0$, and assume for simplicity that there are countably many such values. Appendix A.4 then adapts Heckman and Vytlacil (2001) to show that the LIV estimand in (23) equals a weighted average of MTEs defined with respect to V:

$$LIV(r, p(z, r)) = \sum_{j \in J_{z,r}} w_j(z, r) E[\Delta|R = r, V = v^*_{z,r,j}] \ , \tag{24}$$

[23] As shown by Vytlacil (2002), separability implies the instrument monotoniticy assumption typically invoked in the LATE framework of Imbens and Angrist (1994).

[24] Despite being random quantities, we use lowercase letters to denote the marginal defendant $u^*_{z,r}$ (or $v^*_{z,r}$) below.

where the weights $w_j(z,r)$ add up to one. In the ERM, the weights $w_j(z,r)$ are also non-negative, as additive separability of $\zeta(Z,R,V)$ guarantees that the partial derivatives $\zeta_z(z,r,v^*_{z,r,j})$, which determine the sign of $w_j(z,r)$, must all share the same sign across all marginal values indexed by $j \in J_{z,r}$. More importantly, the ERM ensures that $E[\Delta|R=r,V=v^*_{z,r,j}] = \tau(z,r)$ for all $j \in J_{z,r}$. Putting these results together leads to the following result for the ERM:

$$LIV(r,p(z,r)) = E[\Delta|R=r,U_r=u^*_{z,r}] = E[\Delta|R=r,V=v^*_{z,r,j}] = \tau(z,r) \;\; \forall j \in J_{z,r} \; . \tag{25}$$

In other words, under the ERM, the LIV estimand evaluated at a given z and r identifies the MTE corresponding to judge z's marginal defendants of race r, defined in two ways: in terms of the normalized unobservable U_r typically used in the empirical MTE literature, and also in terms of the unobservable defendant characteristics V from the underlying decision model. These MTEs, in turn, identify the constant value of the benefit function used by judge z when facing defendants of a given race r. The outcome test of τ-bias for a particular judge z can therefore be successfully implemented under the ERM via comparisons of LIV estimands across race, i.e.,

$$LIV(w,p(z,w)) - LIV(b,p(z,b)) = \tau(z,w) - \tau(z,b) \; . \tag{26}$$

Testable Implications of the Extended Roy Model

Our results up to this point have shown that the ERM delivers both a logically valid and econometrically viable outcome test in settings with exogenous and continuous variation across decision makers. The ERM places substantive restrictions on decision maker behavior, however, so researchers interested in imposing the ERM to conduct a valid outcome test must justify its applicability in a given setting. Usefully, the ERM has the following testable implication (not shared by the GRM), which establishes a necessary condition for the suitability of the ERM in a given environment.

Proposition 5.1. Assume the CDF of $E[\Delta|R,V]$ conditional on $R=r$ is one-to-one. Then in the ERM, $E[\Delta|R=r,U_r=p(z,r)]$ is strictly increasing in the propensity score for each $r \in \{w,b\}$.

See Appendix A.3 for the proof. Intuitively, an ERM judge z with a relatively high propensity score for a given race r must be a judge with a relatively high fixed threshold $\tau(z,r)$ of tolerable misconduct effects for that race. Since marginal defendants are those with misconduct effects equal

to this tolerance threshold, the MTEs that identify marginal misconduct effects of defendants of race r facing judge z must be increasing in the judge propensity score p.[25] Since these MTEs are the objects of interest in the outcome test, assessing whether they are strictly increasing—and thus assessing the suitability of the ERM in a given setting—is a natural next step after their estimation.

Remark 5.1. When $E[\Delta|R = r, U_r = p(z, r)]$ is strictly increasing in p for each $r \in \{w, b\}$, the ERM is not rejected by the data, at least with respect to the testable implication in Proposition 5.1. Hull (2021) builds on our results to interpret the modeling restrictions of the ERM as "without loss of generality" when MTEs are strictly increasing.[26] This interpretation confounds the question of whether models are rejected given the data at hand with the question of how to choose between alternative model-based interpretations of a given estimand.

This observation has analogues in many econometric settings. A notable and instructive analogy is the 2SLS estimand with a binary treatment and valid binary instrument, $\frac{Cov(Y,Z)}{Cov(D,Z)}$. This estimand admits at least three alternative interpretations, depending on the underlying model. First, in a model with homogeneous treatment effects across individuals, i.e. $Y_1 - Y_0 = c$ for some constant c, the 2SLS estimand identifies c and thus is interpreted as the constant causal effect of the treatment on the outcome. Alternatively, in a heterogeneous effect model that assumes instrument monotonicity, as in Imbens and Angrist (1994), the 2SLS estimand identifies an average effect of treatment solely among instrument compliers. Finally, in a heterogeneous effects model without monotonicity, the 2SLS estimand identifies a linear combination of average effects among compliers and defiers that generally fails to correspond to an average treatment effect for any (group of) individuals. It is the choice of model—not the data—that ultimately distinguishes between these three alternative and substantively different interpretations of the 2SLS estimand. Stating that the 2SLS estimand admits a homogeneous treatment effect representation among observationally equivalent individuals does not render treatment effects in fact homogeneous. Analogously, failing to reject the ERM based on the testable implication in Proposition 5.1 does not mean imposing the ERM leaves decision maker behavior unrestricted.

[25] This result is similar in spirit to the last sentence of Proposition 3 in Anwar and Fang (2006), which concerns the ranking of average (rather than marginal) search outcomes across discrete police officer races (rather than individual decision makers) who search drivers of a given race at different rates.

[26] Concretely, Hull (2021, p. 12) states that "whenever the race-specific MTE curves are strictly increasing there exists a V such that decisions have an extended Roy model representation."

5.2 Other Settings and Data Environments

In many empirical settings, variation across decision makers Z may only be reasonably characterized as discrete, rather than continuous, and decision makers may not be exogenously assigned to cases. These deviations introduce additional challenges for identifying decision maker bias beyond the logical and econometric issues discussed so far. In this subsection, we discuss these challenges and potential solutions.

Discrete Instruments

We first consider settings in which decision makers Z continue to be exogenously assigned to cases, and behave according to the Extended Roy Model (ERM), but Z may only be reasonably characterized as discrete, rather than continuous. In this case, exploiting variation across different decision makers can nonparametrically identify local average treatment effects (LATEs), but not MTEs. Outcome tests that compare LATEs across groups are more problematic than the MTE-based tests discussed up to this point, as instrument compliers now include individuals beyond those directly on the decision maker's margin of indifference. As such, the inframarginality problem re-emerges as a challenge to identifying bias with discrete instruments (Arnold et al., 2018).

To illustrate this point, consider two judges z and z' such that $p(z,r) > p(z',r)$ for a given race $R = r$. We can write the LATE parameter as

$$LATE^{z}_{z'}(r) = \int_{p(z',r)}^{p(z,r)} \frac{MTE(r,u_r)}{p(z,r) - p(z',r)} du_r \ , \tag{27}$$

such that the set of compliers can be described as $\{u_r : p(z',r) \leq u_r \leq p(z,r)\}$ in terms of U_r or as

$$\mathcal{V}^{z}_{z'}(r) \equiv \{v : \tau(z',r) \leq E[\Delta|r,v] \leq \tau(z,r)\} \tag{28}$$

in terms of the unobservable V entering the decision problem in (4).

A researcher may propose an outcome test in this setting that interprets the difference

$$LATE^{z}_{z'}(w) - LATE^{z}_{z'}(b) \ , \tag{29}$$

as evidence of racial bias. However, even in the context of an ERM where both judges are racially τ-unbiased—i.e., $\tau(z,r) = \tau(z)$ and $\tau(z',r) = \tau(z')$ in (28)—the difference in (29) would not necessarily be equal to zero. This is because the race-specific LATEs in (29) average across inframarginal black and white compliers who differ from the marginal defendants of interest, analogous to the case of average outcome comparisons that we discuss below. Settings where the support of Z is discrete therefore introduce challenges to the outcome test that go beyond those we identified when Z is continuous, and these challenges are similar to those that arise when exogenous instruments are not even available.

Partial Identification

With discrete instruments, instead of attempting to point identify MTEs, one could approach the problem from a different angle and consider partial identification of the MTEs of interest, and bias in turn, by using strategies that restrict the range of admissible parameters using information from the data. Anwar and Fang (2006) offer a pioneering example of this approach: rather than attempting to directly identify the outcomes of marginal drivers of each race, they derive restrictions on the rankings of decision rates and average outcomes across different officer groups under the null hypothesis of unbiased ERM officers.[27] A natural expansion of this approach when Z indexes individual decision makers would apply the methods of Mogstad et al. (2018) to derive bounds on the MTE functions implied by the LATE difference in (29) and other empirical moments that constrain the set of possible MTEs. In the context of the ERM, the identified sets for the MTEs may still be informative enough to make statements about the absence or presence of bias.

Average Outcome Comparisons without Instruments

In many empirical settings, cases are not exogenously assigned to decision makers. In such settings, no valid instrument may be available, and the analyst must use a different identification strategy. A widespread approach in the discrimination literature is to compare *average* post-decision outcomes across groups for a given decision maker (or group of decision makers), typically after conditioning

[27] See also Alesina and La Ferrara (2014) and Marx (2022).

on a set of observable covariates X (which we again suppress for parsimony):[28]

$$E[\Delta|w, z, D = 1] - E[\Delta|b, z, D = 1] \ , \tag{30}$$

where for simplicity we assume the relevant outcome is $\Delta \equiv Y_1$. In pre-trial release, for example, the analyst considers comparing the average rate of pre-trial misconduct among white defendants released by judge z versus the average rate of misconduct among black defendants released by the same judge. The more general case of $\Delta \equiv Y_1 - Y_0$ does not affect the derivations below but brings additional identification challenges, as in some settings the analyst does not observe Y when $D = 0$.

The question of whether the observed difference in average outcomes in (30) is informative about unobserved differences at the margin,

$$E[\Delta|w, v^*_{z,w}] - E[\Delta|b, v^*_{z,b}] \ , \tag{31}$$

is a question about identification. In turn, the question of whether the marginal difference in (31) is informative about decision maker bias is a question about logical validity. From our results in Section 4, we know that unrestricted decision models do not yield logically valid outcome tests of bias as defined in Section 3, even if the marginal outcome difference in (31) is perfectly known.

We therefore examine conditions an analyst could impose not only to render the average outcome difference in (30) informative about the marginal difference in (31), but also to render (31) informative about bias. To study clear and intuitive cases, assume V is scalar, $E[\Delta|r, v]$ is strictly increasing in v, and $\tau(z, r, v)$ is weakly decreasing in v, which ensures a unique type of marginal defendant $v^*_{z,r}$ for each judge z and defendant race r. We also consider all of the analysis below to be conditional on a given judge $Z = z$, which we leave implicit to reduce notation. With these restrictions, the event $D = 1$ is equivalent to $V \leq v^*_{z,r}$; intuitively, if V represents a summary measure of the defendant's criminal history, then judge z releases all defendants of race r with criminal history below the cutoff value $v^*_{z,r}$, and detains all those above. This allows us to write

[28] Peterson (1981) and Berkovec et al. (1994) are early examples of this approach in the context of gender and racial discrimination in lending, discussed in Ross and Yinger (2003). Charles and Guryan (2011) review similar approaches in the context of racial discrimination in the labor market. More recent applications include gender bias in academic publishing (Card and DellaVigna, 2020; Card et al., 2020; Hengel and Moon, 2020) and racial bias in police use of force (Fryer, 2019).

the estimand in (30) as

$$\int E[\Delta|w,v]dF_{V|R=w,V\leq v^*_{z,w}}(v) - \int E[\Delta|b,v]dF_{V|R=b,V\leq v^*_{z,b}}(v) \ . \tag{32}$$

This formulation highlights at least two distinct reasons why average outcome comparisons, without further restrictions, fail to be informative about decision maker bias. Both involve the fact that the two averages in (32) integrate over non-marginal values of V, sweeping in inframarginal individuals who are irrelevant for the marginal comparison in (31)—i.e. the well-known "inframarginality" problem in the outcome test literature (e.g., Heckman et al., 1998; Knowles et al., 2001; Ayres, 2002; Anwar and Fang, 2006). First, the distribution of V generally differs across R. Integrating over non-marginal white and black defendants with arbitrarily different distributions of V therefore allows the magnitude and sign of (32) to be arbitrarily different from that of (31). Second, expected outcomes conditional on race and non-race characteristics, $E[\Delta|R,V]$, also generally vary across R. The two averages in (32) thus integrate arbitrarily different conditional expectation functions over inframarginal values of V, which is another reason the magnitude and sign of (32) can differ arbitrarily from that of (31).

An analyst may attempt to solve the first problem by assuming that V is statistically independent of R (again, typically after conditioning on X and Z, which we keep implicit):[29]

$$V \perp\!\!\!\perp R \ . \tag{33}$$

Imposing (33) alone is insufficient, however, since the second problem still allows (32) to integrate arbitrarily different conditional expectation functions over inframarginal values of V. An interesting takeaway from this point is that assuming away racial differences in the characteristics observed by the decision maker but not the analyst is not a sufficient condition for average outcome comparisons like (30) to be informative about bias.

Likewise, the analyst may attempt to solve the second problem by excluding race R from the

[29]The assumption in (33) does not require specifying the functional form of F_V. See Simoiu et al. (2017) and Arnold et al. (2022) for alternative approaches that allow V to be distributed differently across R but impose parametric assumptions on these distributions.

expected cost/outcome function:

$$E[\Delta|R,V] = E[\Delta|V] \ . \tag{34}$$

As discussed in Remark 4.6 and shown in Appendix C, this exclusion restriction (along with the shape restrictions imposed to derive (32)) is sufficient for a logically valid outcome test, in that the marginal comparison in (31) becomes informative about whether judge z is globally τ-biased. Imposing (34) is not sufficient on its own, however, to render the average difference in (32) informative about the marginal difference in (31), leaving the average outcome difference still uninformative about bias. The reason is that even if the outcome functions $E[\Delta|V]$ being integrated in (32) are the same for white and black defendants with the same V, they are still weighted in the integration by the arbitrarily different distributions $F_{V|R=r,V\leq v^*_{z,r}}$ across race r, as discussed above.

Imposing only one of (33) or (34) in isolation is therefore insufficient to render the average outcome comparison in (32) informative about bias. Imposing both restrictions simultaneously, however, does achieve this goal, at least with regard to global τ-bias and under the additional shape restrictions imposed to derive (32). To see this, write (32) under all of these restrictions as

$$\int E[\Delta|v]dF_{V|V\leq v^*_{z,w}}(v) - \int E[\Delta|v]dF_{V|V\leq v^*_{z,b}}(v) \ . \tag{35}$$

First, suppose judge z is racially τ-unbiased, such that $\tau(z,r,v) = \tau(z,v)$. Since neither the cost function $E[\Delta|V]$ nor the benefit function $\tau(z,v)$ depend on r, neither does their intersection point defining marginal defendants: $v^*_{z,w} = v^*_{z,b} = v^*_z$. Then the average outcome difference in (35) collapses to zero. Now suppose judge z is globally τ-biased against black defendants, such that $\tau(z,w,v) > \tau(z,b,v)$ for all v. Since $E[\Delta|V]$ is strictly increasing in V and $\tau(z,r,v)$ is weakly decreasing in V for each r, it follows that $v^*_{z,w} > v^*_{z,b}$. In addition, with strictly increasing $E[\Delta|v]$, it follows from the properties of truncated distributions (and truncated expectations) that

$$\frac{\partial}{\partial \tilde{v}} \int E[\Delta|v]dF_{V|V\leq \tilde{v}}(v) > 0 \ ,$$

which implies the average outcome difference in (35) is positive given $v^*_{z,w} > v^*_{z,b}$. By an exactly parallel argument, if judge z is globally τ-biased against white defendants, the average outcome difference in (35) is negative.

Thus, if the analyst is willing to impose both (33) and (34) (and the shape restrictions leading to (32)), then the average outcome comparison in (30) will be informative about whether judge z is globally τ-biased. To understand the intuition of this result, observe that (33) and (34) together eliminate any role for race to influence expected outcomes, either directly as a relevant predictor or indirectly through correlations with other non-race characteristics observed by the judge, which eliminates any rationale for statistical discrimination. The only remaining channel through which race could influence judge decisions and the outcomes of released defendants, both marginal and inframarginal, is through the benefit function $\tau(\cdot)$, i.e. through decision maker τ-bias.

Remark 5.2. The result above illustrates conditions under which average outcome comparisons can be informative about the *sign* of judge z's τ-bias. Contrasting the magnitude of (35) across different judges z and z', however, would not be informative about whether z is more or less τ-biased than z', even if judges were additionally assumed to satisfy the ERM and receive randomly assigned cases. This is because, in contrast to the result in (16), the difference in (35) also integrates over inframarginal individuals. Different judges can set different cutoff levels $v^*_{z,b}$, which can lead to arbitrarily different magnitudes of (35) even for ERM judges with the same magnitude of τ-bias, which is only equal to expected outcome differences at the margin.

Remark 5.3. In the case of exogenous and continuous instrumental variation across decision makers studied in Section 5.1, imposing an Extended Roy Model of decision maker behavior was sufficient for a logically valid and econometrically identified outcome test. In contrast, in the case of observational average comparisons like (30), imposing the ERM is neither necessary—as the result above did not impose it—nor sufficient, as the inframarginality issues discussed after (32) would remain regardless of whether an ERM were imposed.

Remark 5.4. Interestingly, imposing both (33) and (34) (and the shape restrictions leading to (35)) to justify an average outcome test would actually render observing outcomes unnecessary. The analyst could conduct an equivalent benchmark test that simply compares decision *rates* across race for a given decision maker,

$$P\{D = 1|R = w, Z = z\} - P\{D = 1|R = b, Z = z\} = F_V(v^*_{z,w}) - F_V(v^*_{z,b}) \ . \tag{36}$$

since the sign of (36) must match the sign of (35) under these restrictions. The intuition here is that while benchmark tests generally fail to be informative about bias because of racial differences in unobservables and statistical discrimination, those are exactly the two phenomena shut down by (33) and (34). With both outcomes and decisions observed, a testable condition of these restrictions is that the sign of the average outcome test in (30) and the benchmark test in (36) must match.

6 Conclusion

In this paper, we have carefully examined what researchers can and cannot learn about bias in decision making from outcome tests. We showed that models of decision making underpinning outcome tests can be usefully recast as Roy models, since heterogeneous potential outcomes enter directly into the decision maker's choice equation. Different members of the Roy model family, however, are distinguished by the tightness of the link between potential outcomes and decisions, and we showed that these distinctions have important implications for defining bias, deriving logically valid outcome tests of such bias, and identifying the marginal outcomes that the test requires. Together, these results offer a methodological guide for researchers considering the use of outcome tests to detect bias across a wide range of empirical environments.

References

Alesina, A. and E. La Ferrara (2014): "A Test of Racial Bias in Capital Sentencing," *American Economic Review*, 104, 3397–3433.

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

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

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

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.

——— (2012): "Testing for the Role of Prejudice in Emergency Departments Using Bounceback Rates," *The B.E. Journal of Economic Analysis & Policy*, 13.

——— (2015): "Testing for Racial Prejudice in the Parole Board Release Process: Theory and Evidence," *The Journal of Legal Studies*, 44, 1–37.

Arcidiacono, P., J. Kinsler, and T. Ransom (2021): "Legacy and Athlete Preferences at Harvard," *Journal of Labor Economics*, 40, 133–156.

Arcidiacono, P. and M. Lovenheim (2016): "Affirmative Action and the Quality-Fit Trade-off," *Journal of Economic Literature*.

Arnold, D., W. Dobbie, and P. Hull (2022): "Measuring Racial Discrimination in Bail Decisions," *American Economic Review*, Forthcoming.

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

——— (2020): "Correction Appendix to 'Racial Bias in Bail Decisions'," *Mimeo*.

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

Ash, E., S. Asher, A. Bhowmick, S. Bhupatiraju, D. Chen, T. Devi, Goessmann Christoph, P. Novosad, and B. Siddiqi (2022): "In-Group Bias in the Indian Judiciary: Evidence from 5 Million Criminal Cases," *Working paper*.

Ayres, I. (2002): "Outcome Tests of Racial Disparities in Police Practices," *Justice Research and Policy*, 4, 131–142.

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.

——— (1993): "Nobel Lecture: The Economic Way of Looking at Behavior," *Journal of Political Economy*, 101, 385–409.

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.

Berkovec, J. A., G. B. Canner, S. A. Gabriel, and T. H. Hannan (1994): "Race, Redlining, and Residential Mortgage Loan Performance," *The Journal of Real Estate Finance and Economics*, 9, 263–294.

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.

Bhattacharya, D. and J. Shvets (2022): "Inferring Trade-Offs in University Admissions: Evidence from Cambridge," *Working paper*.

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

Bjerk, D. (2007): "Racial Profiling, Statistical Discrimination, and the Effect of a Colorblind Policy on the Crime Rate," *Journal of Public Economic Theory*, 9, 521–545.

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.

Bleemer, Z. (2022): "Affirmative Action, Mismatch, and Economic Mobility after California's Proposition 209," *The Quarterly Journal of Economics*, 137, 115–160.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2021): "Inaccurate Statistical Discrimination: An Identification Problem," *Working paper*.

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.

Card, D. and S. DellaVigna (2020): "What Do Editors Maximize? Evidence from Four Economics Journals," *The Review of Economics and Statistics*, 102, 195–217.

Card, D., S. DellaVigna, P. Funk, and N. Iriberri (2020): "Are Referees and Editors in Economics Gender Neutral?" *The Quarterly Journal of Economics*, 135, 269–327.

Charles, K. K. and J. Guryan (2011): "Studying Discrimination: Fundamental Challenges and Recent Progress," *Annual Review of Economics*, 3, 479–511.

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

Dharmapala, D. and S. L. Ross (2004): "Racial Bias in Motor Vehicle Searches: Additional Theory and Evidence," *Contributions in Economic Analysis & Policy*, 3, 1–21.

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 (2021): "Measuring Bias in Consumer Lending," *The Review of Economic Studies*, 88, 2799–2832.

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

Dominitz, J. and J. Knowles (2006): "Crime Minimisation and Racial bias: What Can We Learn from Police Search Data?" *The Economic Journal*, 116, F368–F384.

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.

Fryer, R. G. (2019): "An Empirical Analysis of Racial Differences in Police Use of Force," *Journal of Political Economy*, 127, 1210–1261.

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

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

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 E. Vytlacil (2001): "Local Instrumental Variables," in *Nonlinear Statistical Modeling*, ed. by C. Hsiao, K. Morimune, and J. L. Powell, Cambridge: Cambridge University Press.

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.

Hengel, E. and E. Moon (2020): "Gender and Quality at Top Economics Journals," *Working paper*.

Hernández-Murillo, R. and J. Knowles (2004): "Racial Profiling Or Racist Policing? Bounds Tests In Aggregate Data," *International Economic Review*, 45, 959–989.

Hull, P. (2021): "What Marginal Outcome Tests Can Tell Us About Racially Biased Decision-Making," *Working paper*.

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

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.

Knepper, M. (2018): "When the Shadow Is the Substance: Judge Gender and the Outcomes of Workplace Sex Discrimination Cases," *Journal of Labor Economics*, 36, 623–664.

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

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

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.

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.

Manski, C. F. (2006): "Search Profiling with Partial Knowledge of Deterrence," *The Economic Journal*, 116, F385–F401.

Marx, P. (2022): "An Absolute Test of Racial Prejudice," *The Journal of Law, Economics, and Organization*, 38, 42–91.

Mechoulan, S. and N. Sahuguet (2015): "Assessing Racial Disparities in Parole Release," *The Journal of Legal Studies*, 44, 39–74.

Mogstad, M., A. Santos, and A. Torgovitsky (2018): "Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters," *Econometrica*, 86, 1589–1619.

Mountjoy, J. and B. R. Hickman (2021): "The Returns to College(s): Relative Value-Added and Match Effects in Higher Education," *NBER Working Paper No. 29276*.

Persico, N. (2002): "Racial Profiling, Fairness, and Effectiveness of Policing ," *American Economic Review*, 92, 1472–1497.

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

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

Peterson, R. L. (1981): "An Investigation of Sex Discrimination in Commercial Banks' Direct Consumer Lending," *The Bell Journal of Economics*, 12, 547–561.

Ross, S. L. and J. Yinger (2003): *The Color of Credit : Mortgage Discrimination, Research Methodology, and Fair-Lending Enforcement*, Cambridge, Mass: The MIT Press.

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.

Shayo, M. and A. Zussman (2011): "Judicial Ingroup Bias in the Shadow of Terrorism," *The Quarterly Journal of Economics*, 126, 1447–1484.

Simoiu, C., S. Corbett-Davies, and S. Goel (2017): "The Problem of Infra-Marginality in Outcome Tests for Discrimination," *Annals of Applied Statistics*, 11, 1193–1216.

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

Supplemental Appendix: "On the Use of Outcome Tests for Detecting Bias in Decision Making"

Ivan A. Canay, Magne Mogstad, and Jack Mountjoy

June 10, 2023

A Proofs of the Main Theorems

Throughout the appendix, we define

$$\Lambda(R, V) \equiv E[\Delta | R, V] \ , \tag{A.1}$$

and use the following spaces for the pair of functions $\Lambda(r, \cdot) : \mathcal{V} \to \mathbf{R}$ and $\tau(z, r, \cdot) : \mathcal{V} \to \mathbf{R}$. In addition, to simplify our argument we assume throughout the appendix that V is scalar. We note, however, that our results would extend to the case where V is a vector but assuming that the conditions we require below hold for one of the components in V conditional on the rest of them.

Definition A.1. For fixed values of $r \in \{w, b\}$ and $z \in \mathcal{Z}$, $\mathcal{F}$ denotes the space of all pairs of functions $\Lambda(r, \cdot) : \mathcal{V} \to \mathbf{R}$ and $\tau(z, r, \cdot) : \mathcal{V} \to \mathbf{R}$.

Definition A.2. Let $\mathcal{V}^0$ be an open subset of $\mathcal{V}$. For fixed values of $r \in \{w, b\}$ and $z \in \mathcal{Z}$, $\mathcal{F}^{\mathrm{m}}(\mathcal{V}^0) \subseteq \mathcal{F}$ denotes the space of all pairs of functions $\Lambda(r, \cdot)$ and $\tau(z, r, \cdot)$ such that:

1. $\Lambda(r, \cdot) : \mathcal{V} \to \mathcal{I} \subseteq \mathbf{R}$ is weakly monotone in $v \in \mathcal{V}^0$.
2. $\tau(z, r, \cdot) : \mathcal{V} \to \mathcal{I} \subseteq \mathbf{R}$ is weakly monotone in $v \in \mathcal{V}^0$.
3. For $I_{\mathrm{l}} \equiv \inf \mathcal{I}$ and $I_{\mathrm{u}} \equiv \sup \mathcal{I}$, $I_{\mathrm{l}} < \tau(z, r, v) < I_{\mathrm{u}}$ for all $v \in \mathcal{V}^0$.
4. There exist $v^*_{z,r} \in \mathcal{V}^0$ such that $\Lambda(r, v^*_{z,r}) = \tau(z, r, v^*_{z,r})$.
5. The sets

$$\mathcal{V}^0_{\mathrm{l}} \equiv \{v \in \mathcal{V}^0 : \tau(z, w, v) < \tau(z, w, v^*_{z,w})\} \ , \text{ and} \tag{A.2}$$

$$\mathcal{V}^0_{\mathrm{u}} \equiv \{v \in \mathcal{V}^0 : \tau(z, w, v^*_{z,w}) < \tau(z, w, v)\} \tag{A.3}$$

are non-empty open subsets of $\mathcal{V}^0$.

Definition A.3. For fixed values of $r \in \{w, b\}$ and $z \in \mathcal{Z}$, $\mathcal{F}^{\mathrm{cm}}(\mathcal{V}^0) \subseteq \mathcal{F}^{\mathrm{m}}(\mathcal{V}^0)$ denotes the space of all pairs of functions $\Lambda(r, \cdot)$ and $\tau(z, r, \cdot)$ satisfying the conditions in Definition A.2, where conditions 1 and 2 additionally require the functions to be continuous.

Remark A.1. The space of functions $\mathcal{F}^{\mathrm{m}}(\mathcal{V}^0)$ is quite general and flexible enough to accommodate a variety of cases. To start, the space does not restrict the functions outside $\mathcal{V}^0$, but working with $\mathcal{F}^{\mathrm{m}}(\mathcal{V})$ the restrictions become global. Conditions 1 and 2 impose weak monotonicity. Note that the condition does not state whether the monotonicity should be increasing or decreasing and that it does not impose continuity, thus allowing for step functions. Condition 3 requires τ to be in the interior of its range on $\mathcal{V}^0$ to allow for other functions to be either above or below it. Condition 4 requires crossing on $\mathcal{V}^0$. Condition 5 essentially requires τ to vary on $\mathcal{V}^0$ and have a set of points with "high" values and a set of points with "low" values. This assumption is only required for one value of r and, without loss of generality, here we assume that such value is w.

A.1 Proof of Theorem 4.1

Note that, by definition of $v^*_{z,r}$, the difference in (15) is the same as

$$\tau(z, w, v^*_{z,w}) - \tau(z, b, v^*_{z,b}) \ .$$

The proof for functions in $\mathcal{F}$ is therefore trivial, as these are two different functions evaluated at two different points so we omit a detailed treatment here. The result becomes more interesting as we add more restrictions, starting with $\mathcal{F}^{\mathrm{m}}(\mathcal{V}^0)$. We divide the proof in cases and assume without loss of generality that $\Lambda(r, v)$, as defined in (A.1), is weakly increasing as the other case involves symmetric arguments.

Case (i). When z is racially τ-unbiased the function $\tau(z, r, v)$ does not depend on r and becomes $\tau(z, v)$. Take $(\Lambda(w, v), \tau(z, v)) \in \mathcal{F}^{\mathrm{m}}(\mathcal{V}^0)$ and recall $v^*_{z,w} \in \mathcal{V}^0$ is such that $\Lambda(w, v^*_{z,w}) = \tau(z, v^*_{z,w})$. Choose a point in $\mathcal{V}^0_{\mathrm{u}}$ such that $\tau(z, v)$ is continuous at the point and call it $v^*_{z,b}$. By Definition A.2(5), $\tau(z, v^*_{z,b}) > \tau(z, v^*_{z,w})$. Next, define two functions, $H_r(v)$ and $G_r(v)$, as follows. First,

$$\begin{aligned} H_r(v) = \{&\text{continuous and weakly increasing function mapping } \mathcal{V}^0 \cap \{v \leq v^*_{z,r}\} \text{ to } \mathcal{I} \\ &\text{s.t. } H_r(v) < \tau(z, r, v) \ \forall v < v^*_{z,r} \text{ and } H_r(v^*_{z,r}) = \tau(z, r, v^*_{z,r})\} \ . \end{aligned} \tag{A.4}$$

Such a function exists provided $\tau(z, r, v) > I_l$ for all $v \in \mathcal{V}^0$ and $\tau(z, r, v)$ is continuous at $v^*_{z,r}$. Second,

$$\begin{aligned} G_r(v) = \{&\text{continuous and strictly increasing function mapping } \mathcal{V}^0 \cap \{v > v^*_{z,r}\} \text{ to } \mathcal{I} \\ &\text{s.t. } G_r(v) > \tau(z, r, v) \ \text{ and } \lim_{v \to v^*_{z,r}} G_r(v) = \tau(z, r, v^*_{z,r})\} \ . \end{aligned} \tag{A.5}$$

Such a function exists provided $\tau(z, r, v) < I_u$ for all $v \in \mathcal{V}^0$ and $\tau(z, r, v)$ is continuous at $v^*_{z,r}$. If we now

define

$$\Lambda(b,v) = \begin{cases} H_b(v) & \text{for } v \in \mathcal{V}^0 \cap \{v \leq v^*_{z,b}\} \\ G_b(v) & \text{for } v \in \mathcal{V}^0 \cap \{v > v^*_{z,b}\} \end{cases} , \tag{A.6}$$

then $\Lambda(b,v)$ belongs to $\mathcal{F}^{\mathrm{m}}(\mathcal{V}^0)$ and satisfies $\Lambda(b,v^*_{z,b}) = \tau(z,v^*_{z,b})$. The case leads to (15) being negative. Replacing $v^*_{z,b} \in \mathcal{V}^0_{\mathrm{u}}$ with $v^*_{z,b} \in \mathcal{V}^0_{\mathrm{l}}$ as defined in Definition A.2(5), so that $\tau(z,v^*_{z,b}) < \tau(z,v^*_{z,w})$, and following the same construction, leads to (15) being positive. Setting $v^*_{z,b} = v^*_{z,w}$ or any other value of $v \in \mathcal{V}^0$ such that $\tau(z,v) = \tau(z,v^*_{z,w})$ leads to (15) being zero. In this last case the construction above works provided $\tau(z,v)$ is continuous at $v^*_{z,w}$; otherwise the functions $H_b(v)$ and $G_b(v)$ need to be slightly modified but the arguments are similar and so we omit them here. Note that the exact same construction works on $\mathcal{F}^{\mathrm{cm}}(\mathcal{V}^0)$ since $\Lambda(b,v)$ was chosen to be a continuous function. Also note that extending $\mathcal{V}^0$ to $\mathcal{V}$ does not change any of the arguments either, and so the result holds on $\mathcal{F}^{\mathrm{cm}}(\mathcal{V})$.

Case (ii). We show the case where the judge is globally biased against black defendants as the other case is symmetric. Note that this situation necessarily implies that $\tau(z,w,v) > I_{\mathrm{l}}$ for all $v \in \mathcal{V}$ so we assume this below. Take $(\Lambda(w,v), \tau(z,w,v)) \in \mathcal{F}^{\mathrm{m}}(\mathcal{V}^0)$ and recall $v^*_{z,w} \in \mathcal{V}^0$ is such that $\Lambda(w,v^*_{z,w}) = \tau(z,w,v^*_{z,w})$. Define $\tau(z,b,v)$ on $\mathcal{V}$ as

$$\tau(z,b,v) = \gamma\tau(z,w,v) + (1-\gamma)I_{\mathrm{l}} , \tag{A.7}$$

where $\gamma \in (0,1)$ is defined below. It follows that $\tau(z,b,v) < \tau(z,w,v)$ for all $v \in \mathcal{V}$ and judge z is globally biased against black defendants. By construction, $\tau(z,b,v)$ satisfies Definitions A.2(2) and Definitions A.2(3). In particular, note that $\tau(z,b,v)$ inherits the same monotonicity properties that $\tau(z,w,v)$ has. Next choose a point in $\mathcal{V}^0_{\mathrm{u}}$ such that $\tau(z,w,v)$ is continuous at the point, call it $v^*_{z,b}$, and then choose γ to satisfy

$$\frac{\gamma}{1-\gamma} > \frac{\tau(z,w,v^*_{z,w}) - I_{\mathrm{l}}}{\tau(z,w,v^*_{z,b}) - \tau(z,w,v^*_{z,w})} > 0 , \tag{A.8}$$

where the last inequality follows from Definition A.2(5) implying $\tau(z,w,v^*_{z,b}) > \tau(z,w,v^*_{z,w})$. Finally, define $\Lambda(b,v)$ as in (A.6). Definitions A.2(1) and A.2(4) immediately hold, so we conclude that $(\Lambda(b,v), \tau(z,b,v)) \in \mathcal{F}^{\mathrm{m}}(\mathcal{V}^0)$. Putting all the pieces together,

$$\Lambda(b,v^*_{z,b}) = \tau(z,b,v^*_{z,b}) > \tau(z,w,v^*_{z,w}) = \Lambda(w,v^*_{z,w}) , \tag{A.9}$$

where the strict inequality follows (A.8). The case leads to (15) being negative. For the case that leads to (15) being zero, simply choose γ to satisfy (A.8) with the first inequality replaced by an equality. For the case that leads to (15) being positive, set $v^*_{z,b} = v^*_{z,w}$ and choose any value of $\gamma \in (0,1)$. Note that the exact same construction works on $\mathcal{F}^{\mathrm{cm}}(\mathcal{V}^0)$ since $\Lambda(b,v)$ was chosen to be a continuous function and $\tau(z,b,v)$ in

(A.7) would be continuous on $\mathcal{V}^0$ if $\tau(z,w,v)$ is also continuous. Finally, note that extending $\mathcal{V}^0$ to $\mathcal{V}$ does not change any of the arguments either, and so the result holds on $\mathcal{F}^{\text{cm}}(\mathcal{V})$.

Case (iii). We show the case where the judge is locally biased against black defendants as the other case is symmetric. Note that if $\mathcal{V}_{\text{si}} \cap \mathcal{V}^0 = \emptyset$, then $\tau(z,r,v) = \tau(z,v)$ for all $v \in \mathcal{V}^0$ and the proof follows from the proof of Case (i). We therefore consider a construction where $\mathcal{V}_{\text{si}} \cap \mathcal{V}^0 \neq \emptyset$ and focus on the case where $\mathcal{V}_{\text{si}} = (\underline{v}, \bar{v}) \subseteq \mathcal{V}^0$ since other cases require essentially the same arguments. Take $(\Lambda(w,v), \tau(z,w,v)) \in \mathcal{F}^{\text{m}}(\mathcal{V}^0)$ and recall $v^*_{z,w} \in \mathcal{V}^0$ is such that $\Lambda(w, v^*_{z,w}) = \tau(z,w,v^*_{z,w})$.

Start with the case where $\tau(z,w,v)$ is decreasing and note that it must be case that $\tau(z,w,v) > \tau(z,w,\bar{v})$ for $v < \bar{v}$ in order for $\tau(z,b,v)$ to satisfy Definition 3.4 without violating weak monotonicity at $\bar{v}$. Define $\tau(z,b,v)$ on $\mathcal{V}$ as

$$\tau(z,b,v) \equiv \begin{cases} \gamma\tau(z,w,v) + (1-\gamma)\tau(z,w,\bar{v}) & \text{for } v \in (\underline{v}, \bar{v}) \\ \tau(z,w,v) & \text{for } v \notin (\underline{v}, \bar{v}) \end{cases} , \tag{A.10}$$

for some $\gamma \in (0,1)$. It follows that $\tau(z,b,v) < \tau(z,w,v)$ for all $v \in (\underline{v}, \bar{v})$ and judge z is locally biased against black defendants. By construction, $\tau(z,b,v)$ satisfies Definitions A.2(2) and Definitions A.2(3). To check that $\tau(z,b,v)$ is decreasing, note that for any $v' \leq \underline{v}$ and $v'' \in (\underline{v}, \bar{v})$,

$$\tau(z,b,v') = \tau(z,w,v') > \gamma\tau(z,w,v') + (1-\gamma)\tau(z,w,\bar{v}) \geq \gamma\tau(z,w,v'') + (1-\gamma)\tau(z,w,\bar{v}) = \tau(z,b,v'') ,$$

with a similar argument applying to points $v'' \in (\underline{v}, \bar{v})$ and $v''' \geq \bar{v}$. Next choose a point in $\mathcal{V}^0_{\text{u}}$ such that $\tau(z,w,v)$ is continuous at the point and call it $v^*_{z,b}$. Since $\tau(z,w,v)$ is decreasing, we can choose $v^*_{z,b}$ to satisfy $v^*_{z,b} \leq \underline{v}$. Finally, define $\Lambda(b,v)$ as in (A.6). Definitions A.2(1) and A.2(4) immediately hold, so we conclude that $(\Lambda(b,v), \tau(z,b,v)) \in \mathcal{F}^{\text{m}}(\mathcal{V}^0)$. It then follows that

$$\Lambda(b, v^*_{z,b}) = \tau(z,b,v^*_{z,b}) > \tau(z,w,v^*_{z,w}) = \Lambda(w, v^*_{z,w}) , \tag{A.11}$$

where the strict inequality comes from $\tau(z,b,v^*_{z,b}) = \tau(z,w,v^*_{z,b}) > \tau(z,w,v^*_{z,w})$ since $v^*_{z,b} \leq \underline{v}$. This case leads to (15) being negative. For the case that leads to (15) being zero, simply choose $v^*_{z,b} = v^*_{z,w}$ if $v^*_{z,w} \notin (\underline{v}, \bar{v})$. If $v^*_{z,w} \in (\underline{v}, \bar{v})$ then we need a point $v^*_{z,b} \in (\underline{v}, \bar{v})$ such that $\tau(z,w,v^*_{z,b}) > \tau(z,w,v^*_{z,w})$ in order to set γ to satisfy

$$\frac{\gamma}{1-\gamma} = \frac{\tau(z,w,v^*_{z,w}) - \tau(z,w,\bar{v})}{\tau(z,w,v^*_{z,b}) - \tau(z,w,v^*_{z,w})} .$$

Otherwise exact equality may not arise unless $\tau(z,w,\underline{v}) = \tau(z,w,v^*_{z,w})$. For the case that leads to (15) being positive, choose $v^*_{z,b}$ in $\mathcal{V}^0_{\text{l}}$ so that $v^*_{z,b} \geq \bar{v}$. The construction we just derived is simple, but it does not cover the case where the space is $\mathcal{F}^{\text{cm}}(\mathcal{V}^0)$ since, even when $\tau(z,w,v)$ is everywhere continuous,

$\tau(z,b,v)$ would exhibit a discontinuity at $\underline{v}$ by construction. This, however, can be easily fixed by re-defining $\tau(z,b,v)$ on $(\underline{v},\underline{v}+\epsilon)$ for some $\epsilon>0$ small as the linear function connecting the points $(\underline{v},\tau(z,w,\underline{v}))$ and $(\underline{v}+\epsilon,\gamma\tau(z,w,\underline{v}+\epsilon)+(1-\gamma)\tau(z,w,\bar{v}))$. When $\tau(z,w,v)$ is decreasing and continuous, we can choose ϵ small enough to guarantee that such a line would be strictly below $\tau(z,w,v)$ on $(\underline{v},\underline{v}+\epsilon)$. Finally, note that extending $\mathcal{V}^0$ to $\mathcal{V}$ does not change any of the arguments, and so the result holds on $\mathcal{F}^{\text{cm}}(\mathcal{V})$.

Next consider the case where $\tau(z,w,v)$ is increasing and note that it must be case that $\tau(z,w,v)>\tau(z,w,\underline{v})$ for $v>\underline{v}$ in order for $\tau(z,b,v)$ to satisfy Definition 3.4 without violating weak monotonicity at $\underline{v}$. Define $\tau(z,b,v)$ on $\mathcal{V}$ as

$$\tau(z,b,v)\equiv\begin{cases}\gamma\tau(z,w,v)+(1-\gamma)\tau(z,w,\underline{v}) & \text{for } v\in(\underline{v},\bar{v})\\ \tau(z,w,v) & \text{for } v\notin(\underline{v},\bar{v})\end{cases}, \tag{A.12}$$

for some $\gamma\in(0,1)$. It follows that $\tau(z,b,v)<\tau(z,w,v)$ for all $v\in(\underline{v},\bar{v})$ and judge z is locally biased against black defendants. By construction, $\tau(z,b,v)$ satisfies Definitions A.2(2) and Definitions A.2(3). To check that $\tau(z,b,v)$ is increasing, note that for any $v'\leq\underline{v}$ and $v''\in(\underline{v},\bar{v})$,

$$\tau(z,b,v')\leq\tau(z,w,\underline{v})=\gamma\tau(z,w,\underline{v})+(1-\gamma)\tau(z,w,\underline{v})<\gamma\tau(z,w,v'')+(1-\gamma)\tau(z,w,\underline{v})=\tau(z,b,v'')\ ,$$

with a similar argument applying to points $v''\in(\underline{v},\bar{v})$ and $v'''\geq\bar{v}$. Next choose a point in $\mathcal{V}^0_{\text{u}}$ such that $\tau(z,w,v)$ is continuous at the point and call it $v^*_{z,b}$. Since $\tau(z,w,v)$ is increasing, we can choose $v^*_{z,b}$ to satisfy $v^*_{z,b}\geq\bar{v}$. Finally, define $\Lambda(b,v)$ as in (A.6). Definitions A.2(1) and A.2(4) immediately hold, so we conclude that $(\Lambda(b,v),\tau(z,b,v))\in\mathcal{F}^{\text{m}}(\mathcal{V}^0)$. It then follows that

$$\Lambda(b,v^*_{z,b})=\tau(z,b,v^*_{z,b})>\tau(z,w,v^*_{z,w})=\Lambda(w,v^*_{z,w})\ , \tag{A.13}$$

where the strict inequality comes from $\tau(z,b,v^*_{z,b})=\tau(z,w,v^*_{z,b})>\tau(z,w,v^*_{z,w})$ since $v^*_{z,b}\geq\bar{v}$. This case leads to (15) being negative. For the case that leads to (15) being zero, simply choose $v^*_{z,b}=v^*_{z,w}$ if $v^*_{z,w}\notin(\underline{v},\bar{v})$. If $v^*_{z,w}\in(\underline{v},\bar{v})$ then we need a point $v^*_{z,b}\in(\underline{v},\bar{v})$ such that $\tau(z,w,v^*_{z,b})>\tau(z,w,v^*_{z,w})$ in order to set γ to satisfy

$$\frac{\gamma}{1-\gamma}=\frac{\tau(z,w,v^*_{z,w})-\tau(z,w,\underline{v})}{\tau(z,w,v^*_{z,b})-\tau(z,w,v^*_{z,w})}\ .$$

Otherwise exact equality may not arise unless $\tau(z,w,\bar{v})=\tau(z,w,v^*_{z,w})$. For the case that leads to (15) being positive, choose $v^*_{z,b}$ in $\mathcal{V}^0_{\text{l}}$ so that $v^*_{z,b}\leq\underline{v}$. The construction we just derived is simple, but it does not cover the case where the space is $\mathcal{F}^{\text{cm}}(\mathcal{V}^0)$ since, even when $\tau(z,w,v)$ is everywhere continuous, $\tau(z,b,v)$ would exhibit a discontinuity at $\bar{v}$ by construction. This, however, can be easily fixed by re-defining $\tau(z,b,v)$ on

$(\bar{v}-\epsilon,\bar{v})$ for some $\epsilon>0$ small as the linear function connecting the points $(\bar{v}-\epsilon,\gamma\tau(z,w,\bar{v}-\epsilon)+(1-\gamma)\tau(z,w,\underline{v}))$ and $(\bar{v},\tau(z,w,\bar{v}))$. When $\tau(z,w,v)$ is increasing and continuous, we can choose ϵ small enough to guarantee that such a line would be strictly below $\tau(z,w,v)$ on $(\bar{v}-\epsilon,\bar{v})$. Finally, note that extending $\mathcal{V}^0$ to $\mathcal{V}$ does not change any of the arguments, and so the result holds on $\mathcal{F}^{\text{cm}}(\mathcal{V})$.

Case (iv). This case is similar to the other cases so we only characterize the simplest situation and omit the details of the rest. Take $(\Lambda(w,v),\tau(z,w,v))\in\mathcal{F}^{\text{cm}}(\mathcal{V}^0)$ and recall $v^*_{z,w}\in\mathcal{V}^0$ is such that $\Lambda(w,v^*_{z,w})=\tau(z,w,v^*_{z,w})$. Consider the case where $\tau(z,w,v)$ is strictly decreasing for simplicity and define

$$\tau(z,b,v)\equiv\gamma\tau(z,w,v)+(1-\gamma)\tau(z,w,v^*_{z,w})\ , \tag{A.14}$$

for $\gamma\in(0,1)$. By construction, $\tau(z,b,v)$ satisfies Definitions A.3(2) and Definitions A.3(3). In addition, $\tau(z,b,v)<\tau(z,w,v)$ for $v<v^*_{z,w}$, $\tau(z,b,v^*_{z,w})=\tau(z,w,v^*_{z,w})$, and $\tau(z,b,v)>\tau(z,w,v)$ for $v>v^*_{z,w}$. The judge is then unclassified, as she exhibits higher expected benefits for one race or the other one depending on the values of v. Next, for a given value of $v^*_{z,b}$, define $\Lambda(b,v)$ as in (A.6). Definitions A.3(1) and A.3(4) immediately hold, so we conclude that $(\Lambda(b,v),\tau(z,b,v))\in\mathcal{F}^{\text{cm}}(\mathcal{V}^0)$. To finish the argument, note that by strict monotonicity of $\tau(z,w,v)$,

$$\tau(z,b,v)=\gamma\tau(z,w,v)+(1-\gamma)\tau(z,w,v^*_{z,w})>\gamma\tau(z,w,v^*_{z,w})+(1-\gamma)\tau(z,w,v^*_{z,w})=\tau(z,w,v^*_{z,w}) \tag{A.15}$$

for any $v<v^*_{z,w}$ and

$$\tau(z,b,v)=\gamma\tau(z,w,v)+(1-\gamma)\tau(z,w,v^*_{z,w})<\gamma\tau(z,w,v^*_{z,w})+(1-\gamma)\tau(z,w,v^*_{z,w})=\tau(z,w,v^*_{z,w}) \tag{A.16}$$

for any $v>v^*_{z,w}$. This means that choosing $v^*_{z,b}<v^*_{z,w}$ leads to (15) being negative, choosing $v^*_{z,b}=v^*_{z,w}$ leads to (15) being zero, and choosing $v^*_{z,b}>v^*_{z,w}$ leads to (15) being positive. This completes the proof.

A.2 Proof of Theorem 4.2

By the definition of the marginal defendant in (12), it must be the case that

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

It follows immediately that

$$\Lambda(w,v^*_{z,w})-\Lambda(b,v^*_{z,b})=\tau(z,w)-\tau(z,b)\ , \tag{A.17}$$

and so the outcome test concludes there is no evidence of racial bias if and only if judge $z \in \mathcal{Z}$ is racially τ-unbiased, i.e., $\tau(z,r) = \tau(z)$. Similarly, the outcome test concludes that judge z is biased against black defendants, $\Lambda(w, v^*_{z,w}) > \Lambda(b, v^*_{z,b})$, if and only if judge $z \in \mathcal{Z}$ is racially τ-biased against black defendants. The same holds for τ-biased against white defendants and this concludes the proof.

A.3 Proof of Proposition 5.1

Let $F_{\Lambda|R}(\cdot)$ denote the CDF of $\Lambda(R,V) \equiv E[\Delta|R,V]$ conditional on R, which is assumed to be one-to-one for each value in the support of R. Evaluating $MTE(r,u)$ at $u = p(z,r)$ yields

$$\begin{aligned} MTE(r,p(z,r)) &= E[\Delta|R=r, U_r = p(z,r)] \\ &= E[\Delta|R=r, F_{\Lambda|r}(\Lambda(r,V)) = F_{\Lambda|r}(\tau(z,r))] \\ &= E[\Delta|R=r, \Lambda(r,V) = \tau(z,r)] \\ &= \tau(z,r) \\ &= F^{-1}_{\Lambda|r}(p(z,r)) \ , \end{aligned} \tag{A.18}$$

where the first and second equality use the representation $D = I\{U_r \leq p(z,r)\}$, the second equality exploits that the event $\{U_r = p(z,r)\}$ is equivalent to the event $\{F_{\Lambda|r}(\Lambda(r,V)) = F_{\Lambda|r}(\tau(z,r))\}$ in the ERM, and the third and last equality follow from $F_{\Lambda|r}(\cdot)$ being one-to-one.

A.4 Interpretation of the LIV Estimand without Separability

Consider the latent index model in (20) without additive separability. For simplicity of the argument, assume the function $\zeta(z,r,v)$ has a countable number of crossing points v such that $\zeta(z,r,v) = 0$ for each (z,r) as in Heckman and Vytlacil (2001, Appendix). The results below, however, conceptually extend to the case where there are unaccountably many crossing points at the cost of more cumbersome notation. Assume additionally that $\zeta(z,r,v)$ is continuously differentiable in (z,v) for each value of r. Let $j \in J_{z,r}$ index the set of v points such that the latent index is zero, so that $\zeta(z,r,v^*_j) = 0$ for all $j \in J_{z,r}$. $|J_{z,r}| > 1$ could arise, for example, when v is a scalar and the function $\tau(z,r,v)$ is non-monotonic in v, or when v is multi-dimensional. Also note that the values $\{v^*_j : j \in J_{z,r}\}$ depend on z and r, but we suppress such dependence throughout this

appendix for parsimony. It follows from standard manipulations (see Section A.4.1 below) that

$$\frac{\partial}{\partial z}E[Y|Z=z,R=r] = -\sum_{j\in J_{z,r}} \frac{\zeta_z(z,r,v_j^*)}{|\zeta_v(z,r,v_j^*)|} f_{v|r}(v_j^*) E[\Delta|R=r,V=v_j^*] \ ,$$
$$\frac{\partial}{\partial z}p(z,r) = -\sum_{j\in J_{z,r}} \frac{\zeta_z(z,r,v_j^*)}{|\zeta_v(z,r,v_j^*)|} f_{v|r}(v_j^*) \ ,$$

where $\zeta_z(\cdot)$ and $\zeta_v(\cdot)$ denote the partial derivatives of $\zeta(\cdot)$ with respect to z and v, and $f_{v|r}(v_j^*)$ is the conditional density of V given $(R=r, Z=z)$, which only depends on $R=r$ since V is independent of Z. The LIV estimand is the ratio of these two terms:

$$LIV(r,p(z,r)) = \frac{\frac{\partial}{\partial z}E[Y|Z=z,R=r]}{\frac{\partial}{\partial z}p(z,r)} = \sum_{j\in J_{z,r}} w_j(z,r) E[\Delta|R=r,V=v_j^*] \ , \tag{A.19}$$

where the weights $w_j(z,r)$ are given by

$$w_j(z,r) = \frac{\frac{\zeta_z(z,r,v_j^*)}{|\zeta_v(z,r,v_j^*)|} f_{v|r}(v_j^*)}{\sum_{j\in J_{z,r}} \frac{\zeta_z(z,r,v_j^*)}{|\zeta_v(z,r,v_j^*)|} f_{v|r}(v_j^*)} \ . \tag{A.20}$$

It follows that the LIV estimand, evaluated at a particular judge z and defendant race r, is a weighted average of $E[\Delta|r,v_j^*]$ associated with each of the crossing points v_j^* for $j \in J_{z,r}$, i.e. all defendants of race r who are marginal for judge z. While these weights necessarily add up to one, they could be negative, meaning that the LIV estimand may be negative even when $E[\Delta|r,v_j^*] > 0$ for all $j \in J_{z,r}$. This is because $\zeta_z(z,r,v_j^*)$ could be positive for some j and negative for others; intuitively, reassigning one of judge z's marginal defendants to a nearby judge z' could raise the value of the benefit function $\tau(\cdot)$, while the same reassignment of another of judge z's marginal defendants could lower the value of the benefit function.

As an illustration, consider a case where $|J_{z,r}| = 2$, so there are two values of V that characterize marginal defendants. This could arise when v is a scalar and the function $\tau(\cdot)$ is quadratic in v, or when v is multi-dimensional. For example, suppose that $v = (v_1, v_2)$ where v_1 takes values $\{0,1\}$ and v_2 takes values in $\mathbf{R}$. Consider the case where $\tau(z,r,v) = \tau_1(z,r) + av_1 - v_2$ for some function $\tau_1(z,r)$ and constant $a \in \mathbf{R}$. Then, for any τ^* in the range of $\tau(\cdot)$ we obtain that $\tau(z,r,v) = \tau^*$ for $v = (1, \tau_1(z,r) + a - \tau^*)$ and $v = (0, \tau_1(z,r) - \tau^*)$.

Additive separability of the latent index $\zeta(z,r,v)$ in z and v guarantees that the weights $w_j(z,r)$ in (A.20) are positive, since in that case $\zeta_z(z,r,v)$ is either positive or negative for all $v \in \mathcal{V}$. Indeed, additive

separability ensures that the derivative $\zeta_z(z, r, v_j^*)$ in (A.20) does not depend on v_j^* and so

$$w_j(z,r) = \frac{\frac{f_{v|r}(v_j^*)}{|\zeta_v(z,r,v_j^*)|}}{\sum_{j \in J_{z,r}} \frac{f_{v|r}(v_j^*)}{|\zeta_v(z,r,v_j^*)|}} > 0 \ . \tag{A.21}$$

A.4.1 Proof of (A.19)

We start with the denominator. In order to simplify the expressions, we remove the dependence of $J_{z,r}$ on (z, r) and simply write J from here on. We also keep implicit regularity assumptions that guarantee that all denominators are non-zero as well as conditions on the conditional (on R) density of V that guarantee that the expressions for the integrals we write below are valid. Let $\mathcal{V}_{z,r} \equiv \{v \in \mathcal{V} : \zeta(z, r, v) \leq 0\}$ and note that we can order the crossing points $\{v_j^* : j \in J\}$ from smallest to largest. By the implicit function theorem,

$$\frac{dv_j^*}{dz} = -\frac{\zeta_z(z, r, v_j^*)}{\zeta_v(z, r, v_j^*)} \tag{A.22}$$

where ζ_z and ζ_v denote the partial derivatives of ζ with respect to z and v. Next, let $f_{v|r}(v)$ be the conditional on (z, r) density of v. Since $Z \perp V$, this conditional density does not depend on z. Finally, consider the case where $\zeta(z, r, v) > 0$ for $v \in (-\infty, v_1^*)$, as the other case involves similar arguments. We can then write

$$\begin{aligned} P\{D = 1|Z = z, R = r\} &= \int_{\mathcal{V}_{z,r}} f_{v|r}(v)dv \\ &= \int_{v_1^*}^{v_2^*} f_{v|r}(v)dv + \int_{v_3^*}^{v_4^*} f_{v|r}(v)dv + \int_{v_5^*}^{v_6^*} f_{v|r}(v)dv + \cdots \ . \end{aligned}$$

Taking derivatives, using (A.22), and noting that $f_{v|r}(v)$ is not a function of z, we get

$$\begin{aligned} \frac{\partial}{\partial z} P\{D = 1|Z = z, R = r\} &= f_{v|r}(v_2^*)\frac{dv_2^*}{dz} - f_{v|r}(v_1^*)\frac{dv_1^*}{dz} + f_{v|r}(v_4^*)\frac{dv_4^*}{dz} - f_{v|r}(v_3^*)\frac{dv_3^*}{dz} \\ &\quad + f_{v|r}(v_6^*)\frac{dv_6^*}{dz} - f_{v|r}(v_5^*)\frac{dv_5^*}{dz} + f_{v|r}(v_8^*)\frac{dv_8^*}{dz} - f_{v|r}(v_7^*)\frac{dv_7^*}{dz} + \cdots \\ &= -f_{v|r}(v_1^*)\frac{\zeta_z(z, r, v_1^*)}{|\zeta_v(z, r, v_1^*)|} - f_{v|r}(v_2^*)\frac{\zeta_z(z, r, v_2^*)}{|\zeta_v(z, r, v_2^*)|} - f_{v|r}(v_3^*)\frac{\zeta_z(z, r, v_3^*)}{|\zeta_v(z, r, v_3^*)|} - \cdots \\ &= -\sum_{j \in J} f_{v|r}(v_j^*)\frac{\zeta_z(z, r, v_j^*)}{|\zeta_v(z, r, v_j^*)|} \ , \end{aligned} \tag{A.23}$$

where the second equality follows because $\zeta_v(z, r, v_j^*) < 0$ for all odd values of j and $\zeta_v(z, r, v_j^*) > 0$ for all even values of j. In the case where $\zeta(z, r, v) < 0$ for $v \in (-\infty, v_1^*)$, the first integral in the expression goes from $-\infty$ to v_1^* and in that case $\zeta_v(z, r, v_j^*) > 0$ for all odd values of j and $\zeta_v(z, r, v_j^*) < 0$ for all even values of j. The resulting expression is the same.

Now consider the numerator. Let $\mathcal{V}_{z,r}$ be defined as before and note that we can order the crossing

points $\{v_j^* : j \in J\}$ from smallest to largest. Consider the case where $\zeta(z,r,v) > 0$ for $v \in (-\infty, v_1^*)$, as the other case involves similar arguments. We can then write

$$\begin{aligned}
E[Y|Z=z,R=r] &= E\left[E[Y|Z=z,R=r,D]|Z=z,R=r\right] \\
&= E[Y_1|Z=z,R=r,D=1]p(z,r) + E[Y_0|Z=z,R=r,D=0](1-p(z,r)) \\
&= E[Y_1|Z=z,R=r,v\in\mathcal{V}_{z,r}]p(z,r) + E[Y_0|Z=z,R=r,v\in\mathcal{V}_{z,r}^c](1-p(z,r)) \\
&= \int_{\mathcal{V}_{z,r}} E[Y_1|Z=z,R=r,V=v]p(z,r)f_{v|r}(v|v\in\mathcal{V}_{z,r}) \\
&+ \int_{\mathcal{V}_{z,r}^c} E[Y_0|Z=z,R=r,V=v](1-p(z,r))f_{v|r}(v|v\in\mathcal{V}_{z,r}^c) \\
&= \int_{\mathcal{V}_{z,r}} E[Y_1|R=r,V=v]f_{v|r}(v)I\{v\in\mathcal{V}_{z,r}\} \qquad \text{(A.24)} \\
&+ \int_{\mathcal{V}_{z,r}^c} E[Y_0|R=r,V=v]f_{v|r}(v)I\{v\in\mathcal{V}_{z,r}^c\}\ , \qquad \text{(A.25)}
\end{aligned}$$

where the first equality follows by the LIE, the second equality follows by the definition of propensity score, and the last equality follows from the definition of the condition density

$$f_{v|r}(v|v\in A) = \frac{f_{v|r}(v)I\{v\in A\}}{\Pr\{v\in A|R=r\}}\ , \qquad \text{(A.26)}$$

and the independence of Z. The term in (A.24) can be worked out as in the previous proof by first writing,

$$\begin{aligned}
\int_{\mathcal{V}_{z,r}} E[Y_1|R=r,V=v]f_{v|r}(v) &= \int_{v_1^*}^{v_2^*} E[Y_1|R=r,V=v]f_{v|r}(v)dv \\
&+ \int_{v_3^*}^{v_4^*} E[Y_1|R=r,V=v]f_{v|r}(v)dv + \int_{v_5^*}^{v_6^*} E[Y_1|R=r,V=v]f_{v|r}(v)dv + \cdots\ .
\end{aligned}$$

and then taking derivatives with respect to z,

$$\begin{aligned}
\frac{\partial}{\partial z}\int_{\mathcal{V}_{z,r}} E[Y_1|R=r,V=v]f_{v|r}(v) &= E[Y_1|R=r,V=v_2^*]f_{v|r}(v_2^*)\frac{dv_2^*}{dz} - E[Y_1|R=r,V=v_1^*]f_{v|r}(v_1^*)\frac{dv_1^*}{dz} \\
&+ E[Y_1|R=r,V=v_4^*]f_{v|r}(v_4^*)\frac{dv_4^*}{dz} - E[Y_1|R=r,V=v_3^*]f_{v|r}(v_3^*)\frac{dv_3^*}{dz} + \cdots \\
&= -E[Y_1|R=r,V=v_1^*]f_{v|r}(v_1^*)\frac{\zeta_z(z,r,v_1^*)}{|\zeta_v(z,r,v_1^*)|} \\
&- E[Y_1|R=r,V=v_2^*]f_{v|r}(v_2^*)\frac{\zeta_z(z,r,v_2^*)}{|\zeta_v(z,r,v_2^*)|} \\
&- E[Y_1|R=r,V=v_3^*]f_{v|r}(v_3^*)\frac{\zeta_z(z,r,v_3^*)}{|\zeta_v(z,r,v_3^*)|} - \cdots \\
&= -\sum_{j\in J} E[Y_1|R=r,V=v_j^*]f_{v|r}(v_j^*)\frac{\zeta_z(z,r,v_j^*)}{|\zeta_v(z,r,v_j^*)|}\ ,
\end{aligned}$$

where we again used that $\zeta_v(z,r,v_j^*) < 0$ for all odd values of j and $\zeta_v(z,r,v_j^*) > 0$ for all even values of j. The term in (A.25) can be worked out similarly since,

$$\int_{\mathcal{V}^c_{z,r}} E[Y_0|R=r,V=v]f_{v|r}(v) = \int_{-\infty}^{v_1^*} E[Y_0|R=r,V=v]f_{v|r}(v)dv + \int_{v_2^*}^{v_3^*} E[Y_0|R=r,V=v]f_{v|r}(v)dv + \int_{v_4^*}^{v_5^*} E[Y_0|R=r,V=v]f_{v|r}(v)dv + \cdots \ .$$

and then taking derivatives with respect to z,

$$\begin{aligned}
\frac{\partial}{\partial z}\int_{\mathcal{V}^c_{z,r}} E[Y_0|R=r,V=v]f_{v|r}(v) &= E[Y_0|R=r,V=v_1^*]f_{v|r}(v_1^*)\frac{dv_1^*}{dz} - E[Y_0|R=r,V=v_2^*]f_{v|r}(v_2^*)\frac{dv_2^*}{dz} \\
&+ E[Y_0|R=r,V=v_3^*]f_{v|r}(v_3^*)\frac{dv_3^*}{dz} - E[Y_0|R=r,V=v_4^*]f_{v|r}(v_4^*)\frac{dv_4^*}{dz} + \cdots \\
&= E[Y_0|R=r,V=v_1^*]f_{v|r}(v_1^*)\frac{\zeta_z(z,r,v_1^*)}{|\zeta_v(z,r,v_1^*)|} + E[Y_0|R=r,V=v_2^*]f_{v|r}(v_2^*)\frac{\zeta_z(z,r,v_2^*)}{|\zeta_v(z,r,v_2^*)|} \\
&+ E[Y_0|R=r,V=v_3^*]f_{v|r}(v_3^*)\frac{\zeta_z(z,r,v_3^*)}{|\zeta_v(z,r,v_3^*)|} + E[Y_0|R=r,V=v_4^*]f_{v|r}(v_4^*)\frac{\zeta_z(z,r,v_4^*)}{|\zeta_v(z,r,v_4^*)|} + \cdots \\
&= \sum_{j\in J} E[Y_0|R=r,V=v_j^*]f_{v|r}(v_j^*)\frac{\zeta_z(z,r,v_j^*)}{|\zeta_v(z,r,v_j^*)|} \ ,
\end{aligned}$$

where we again used (A.22) and the fact that $\zeta_v(z,r,v_j^*) < 0$ for all odd values of j and $\zeta_v(z,r,v_j^*) > 0$ for all even values of j. The expression in (A.19) follows by taking the ratio of these quantities.

B Local IV Fails to Implement the Outcome Test in the GRM

Under the GRM, unlike the ERM, local IV fails to identify the MTEs required to implement the outcome test for two reasons. The first reason is the lack of separability of the latent index in (20), which simply equals $\zeta(z,r,v) \equiv E[\Delta|r,v] - \tau(z,r,v)$ in the GRM and is not necessarily separable in the instrument Z and the unobserved heterogeneity V without further assumptions. Without separability, representing the decision model in terms of a uniformly distributed random variable and a propensity score, as in (22), does not lead to a useful formulation. This may not be obvious, as one could start from the unrestricted index model $D = I\Big\{\zeta(Z,R,V) \leq 0\Big\}$, as in Hull (2021), then let $\tilde{F}_{z,r}(\cdot)$ be the conditional distribution of $\zeta(Z,R,V)$ given $(Z=z,R=r)$, and then write

$$D = I\Big\{\tilde{F}_{Z,R}(\zeta(Z,R,V)) \leq \tilde{F}_{Z,R}(0)\Big\} = I\Big\{\tilde{U}_{Z,R} \leq p(Z,R)\Big\} \text{ with } \tilde{U}_{Z,R}|(Z,R) \sim U[0,1] \ .$$

One could then define an MTE as $E[\Delta|R=r,\tilde{U}_{z,r}=u]$. The problem with this approach is that $\tilde{U}_{Z,R}$ is not independent of Z, even if V is independent of Z. Furthermore, the results in Appendix A.4 show

that without separability the LIV estimand identifies a weighted average of MTEs (see (A.19)), with the caveat that such weights could potentially be negative. This means that under the GRM, the LIV estimand generally fails to identify even a convex combination of MTEs.

The second reason is unrelated to the non-separability of the latent index and holds even if one were to further impose additive separability. Concretely, whenever $\tau(z, r, v)$ is a non-trivial function of v (even if this function were to be assumed separable in z and v), the possibly multiple values of v such that $\zeta(z, r, v) = 0$ do not necesarily share the same value of $E[\Delta|r, v]$. In other words, the GRM does not lead to $E[\Delta|R = r, V = v^*_{z,r,j}]$ being equal across $j \in J_{z,r}$ even if one were to impose the additional assumption that $\zeta(z, r, v)$ is separable in z and v. Imposing additive separability is enough to guarantee that the weights in (A.19) are all non-negative, as shown in Appendix A.4, but it does not pin down a unique value of $E[\Delta|r, v]$ for marginal defendants. As a result, the LIV estimand would identify a convex combination of MTEs across marginal defendants of race r facing judge z who have different values of v, making it difficult to interpret comparisons of such weighted averages across race.

To address these two problems, an analyst may be tempted to impose the following assumptions on the GRM, in hopes of securing identification of MTEs for marginal defendants while otherwise maintaining the GRM as a highly flexible model of decision maker behavior:

Assumption B.1. The latent index in (20) satisfies the following two conditions:

(a) **Additive separability**: $\zeta(z, r, v) = \zeta_1(r, v) - \zeta_2(z, r)$.

(b) **Single crossing**: for each (z, r), there is a unique $v^*_{z,r}$ such that $\zeta(z, r, v^*_{z,r}) = 0$.

Assumption B.1(a), which is equivalent to assuming $\tau(z, r, v) = \tau_1(r, v) + \tau_2(z, r)$, guarantees that the GRM admits the representation in (22). Assumption B.1(b), in turn, guarantees that the marginal defendant for each judge z and race r is uniquely determined. Thus, under the GRM with the additional restrictions of Assumption B.1, the LIV estimand evaluated at a given judge z and defendant race r identifies the unique value of $E[\Delta|r, v^*_{z,r}]$. Importantly, however, these conditions to secure identification of $E[\Delta|r, v^*_{z,r}]$ do not affect the results in Theorem 4.1, which took this parameter as known, and thus do not rectify the logical invalidity of the outcome test under the GRM. This is because additive separability and single-crossing do not remove the dependence of the benefit function $\tau(z, r, v)$ on v, which is the key driver of logical invalidity in the GRM. Thus, an analyst employing a GRM under Assumption B.1 may successfully identify marginal outcomes across groups that are nevertheless uninformative about decision maker bias defined in Section 3.

C Excluding Race from the Cost Function

In Section 4 we showed that marginal-based outcome tests are logically invalid in the GRM and logically valid in the ERM. The ERM differs from the GRM in that it excludes non-race defendant characteristics v from the benefit function $\tau(\cdot)$. However, a close inspection to the intuition provided in Figures 1a and 1b points to another possible restriction in the GRM that would break the arguments behind the proof of Theorem 4.1: excluding race from the cost function,

$$\Lambda(r, v) = \Lambda(v) \text{ for all } v \in \mathcal{V} \ . \tag{A.27}$$

This restriction is not strong enough by itself to recover logical validity of the marginal outcome test. A set of apparent minimal conditions would require both $\Lambda(r, v)$ and $\tau(z, r, v)$ to be monotone and also require assuming away the possibility that judge z is locally racially biased or, what we previously label as, unclassified. The combination of these conditions are enough to make the marginal outcome test valid in the GRM as the following Lemma shows.

Lemma C.1. Assume the following conditions:

1. The expected cost function satisfies (A.27) and is strictly increasing on $\mathcal{V}$.
2. The expected benefit function $\tau(z, r, v)$ is weakly decreasing on $\mathcal{V}$.
3. Judge z is either racially τ-unbiased so that $\tau(z, b, v) = \tau(z, w, v)$ for all $v \in \mathcal{V}$ or is globally τ-biased against race r so that $\tau(z, r', v) > \tau(z, r, v)$ for all $v \in \mathcal{V}$.

It follows that

$$\Lambda(v^*_{z,w}) = \Lambda(v^*_{z,b}) \quad \text{if and only if } z \text{ is racially } \tau\text{-unbiased} \ , \tag{A.28}$$

$$\Lambda(v^*_{z,r'}) > \Lambda(v^*_{z,r}) \quad \text{if and only if } z \text{ is globally } \tau\text{-biased against defendants of race } r \ . \tag{A.29}$$

Proof of Lemma C.1. First, suppose that judge $z \in \mathcal{Z}$ is racially τ-unbiased, i.e., $\tau(z, r, v) = \tau(z, v)$ for all $v \in \mathcal{V}$. In this case neither the expected cost nor the expected benefit depend on race, and since they cross only once by the monotonicity assumptions, it follows that $\tau(z, v^*_z) = \Lambda(v^*_z)$ and $v^*_z = v^*_{z,b} = v^*_{z,w}$. The outcome test immediately concludes absence of racial bias in this case. Next, suppose that $\Lambda(v^*_{z,w}) = \Lambda(v^*_{z,b})$. Since $\Lambda(v)$ is strictly increasing, this implies that $v^*_{z,b} = v^*_{z,w}$ and so $\tau(z, b, v^*_{z,b}) = \tau(z, w, v^*_{z,w})$. Under the assumption that z is either racially τ-unbiased or is globally racially τ-biased, this immediately implies that z is racially unbiased.

Second, suppose that judge z is globally τ-biased against black defendants, i.e.

$$\tau(z, w, v) > \tau(z, b, v) \text{ for all } v \in \mathcal{V} \ . \tag{A.30}$$

The other direction is symmetric. Now split the argument in three cases. First, suppose that $v^*_{z,w} = v^*_{z,b}$ and consider the following argument,

$$\Lambda(v^*_{z,w}) = \tau(z, w, v^*_{z,w}) > \tau(z, b, v^*_{z,w}) = \tau(z, b, v^*_{z,b}) = \Lambda(v^*_{z,b}) = \Lambda(v^*_{z,w}) \ , \tag{A.31}$$

where the equalities follow from the definition of marginal defendants and the inequality follows from (A.30). This is a contradiction. Second, suppose that $v^*_{z,w} < v^*_{z,b}$ and consider the following argument,

$$\Lambda(v^*_{z,w}) = \tau(z, w, v^*_{z,w}) > \tau(z, b, v^*_{z,w}) \geq \tau(z, b, v^*_{z,b}) = \Lambda(v^*_{z,b}) \ , \tag{A.32}$$

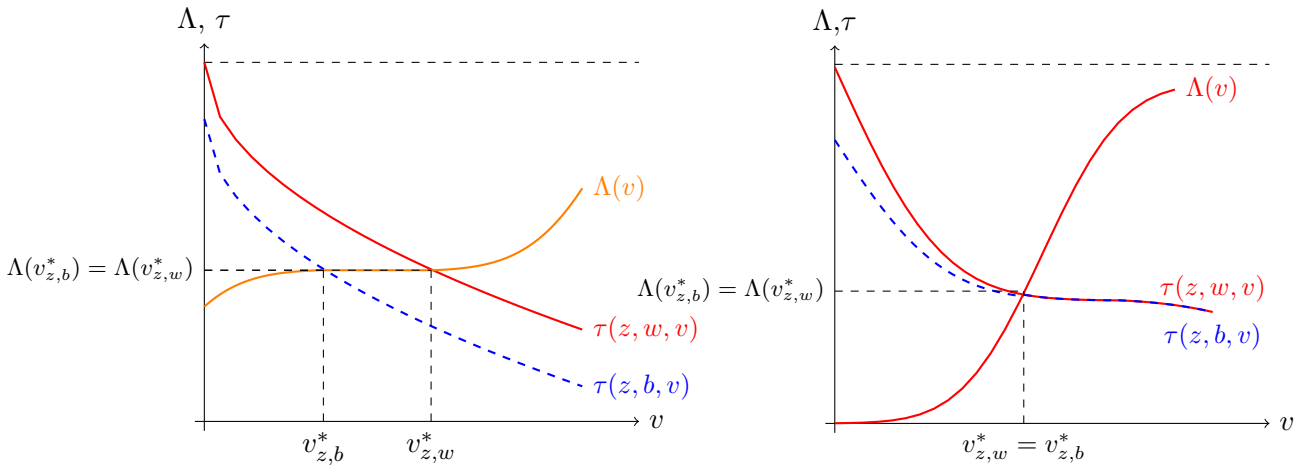
where the equalities follow from the definition of marginal defendants, the first inequality follows from (A.30), and the weak inequality follows from $\tau(z, r, v)$ being weakly decreasing. This leads to a contradiction of $\Lambda(v)$ being strictly increasing. Finally, suppose then that $v^*_{z,w} > v^*_{z,b}$. 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 marginal-based outcome test concludes there is racial bias in the right direction.

To conclude the proof, suppose now that $\Lambda(v^*_{z,w}) > \Lambda(v^*_{z,b})$, so that the outcome test concludes that judge z is biased against black defendants, and assume that $\tau(z, b, v) \geq \tau(z, w, v)$ for all $v \in \mathcal{V}$. By $\Lambda(v)$ being strictly increasing we know that $v^*_{z,w} > v^*_{z,b}$ and so

$$\Lambda(v^*_{z,w}) = \tau(z, w, v^*_{z,w}) \leq \tau(z, w, v^*_{z,b}) \leq \tau(z, b, v^*_{z,b}) = \Lambda(v^*_{z,b}) \ , \tag{A.33}$$

where the equalities follow from the definition of marginal defendants, the first weak inequality follows from $\tau(\cdot)$ being weakly decreasing, and the second weak inequality follows from $\tau(z, b, v) \geq \tau(z, w, v)$ for all $v \in \mathcal{V}$. This leads to a contradiction and completes the proof. ■

Lemma C.1 requires three conditions. Figure A.1a illustrates the strict monotonicity of $\Lambda(v)$ is needed to avoid a situation where the outcome test concludes absence of racial bias simply because $\Lambda(v)$ happens to be flat in the relevant crossing areas. That is, excluding race from the cost function and weak monotonicity of $\Lambda(v)$ are not enough to prevent the outcome test from incorrectly concluding absence of racial bias. The second condition guarantees that there is a unique marginal defendant for each race. If the function $\tau(z, r, v)$ is allowed to be increasing in some subset of $\mathcal{V}$, then the marginal defendants for each race may not be



(a) Weakly Monotone Expected Cost

(b) Locally Racially Bias

Figure A.1: Intuition behind the conditions in Lemma C.1

associated with a unique value of v and the arguments in behind (A.32) and (A.33) would break. Finally, the third condition assumes away the possibility that z is only locally biased against a given race. This is required as, otherwise, a situation could arise where $\Lambda(v^*_{z,w}) = \Lambda(v^*_{z,b})$, the test concludes absence of bias, and despite the fact that $\tau(z, b, v^*_{z,b}) = \tau(z, w, v^*_{z,w})$, the judge is locally racially biased. This is illustrated in Figure A.1b. In that figure, the outcome test concludes that there is no evidence of racial bias when judge z is locally racially biased against black defendants simply because the intersection happens at a point where the expected benefits across race are the same.

D Interpreting (18) Through the Lens of the GRM

Consider the GRM in Definition 2.1 and assume for simplicity that $V \in \mathbf{R}$. Note that $E[\Delta|R, V]$ is a function of (R, V) so the actual treatment effect Δ does not enter D. Then,

$$\begin{aligned} E[D|Z = z, R = w, \Delta = \delta] &= P\left\{E[\Delta|R, V] \leq \tau(Z, R, V)|Z = z, R = w, \Delta = \delta\right\} \\ &= P\left\{\psi_w(V) \leq g_{z,w}(V)|R = w, \Delta = \delta\right\} \ , \end{aligned} \tag{A.34}$$

where the last line used that $V \perp Z$ and changes notation simply to emphasize that these are two functions of V indexed by values of (z, r). Arnold et al. (2022) test for racial discrimination based on

$$T = E\left[P\left\{\psi_w(V) \leq g_{z,w}(V)|R = w, \Delta\right\} - P\left\{\psi_b(V) \leq g_{z,b}(V)|R = b, \Delta\right\}\right] \ . \tag{A.35}$$

To interpret this test in the context of the GRM, one would need to make statements about the distribution of $V|(R,\Delta)$ for the expression inside the expectation, and then about the distribution of Δ for the outer expectation. Without restrictions on these distributions, the values that T can take are not informative about differences in $\tau(Z,R,V)$, even within the context of the ERM. As a simple illustration, consider the following example where judge z is racially τ-unbiased:

$$\psi_w(V) = a_w V \quad \psi_b(V) = a_b V \quad \tau(z,r,V) = -V \quad V|(R,\Delta) \sim U[-\Delta, 1+c_r] \quad \Delta \in \{0,1\} \tag{A.36}$$

where $a_r \in (0,1)$, $c_b > c_w = 0$ and $P\{\Delta = 1\} = 0.5$. It follows that

$$P\Big\{(1+a_w)V \geq 0 | R = w, \Delta = 0\Big\} - P\Big\{(1+a_b)V \geq 0 | R = b, \Delta = 0\Big\} = 1 - 1 = 0 \tag{A.37}$$

$$P\Big\{(1+a_w)V \geq 0 | R = w, \Delta = 1\Big\} - P\Big\{(1+a_b)V \geq 0 | R = b, \Delta = 1\Big\} = \frac{1}{2} - \frac{1+c_b}{2+c_b} < 0 \tag{A.38}$$

since $V|R = b, \Delta = 1 \sim U[-1, 1+c_b]$ and so $P\left\{(1+a_b)V \geq 0 | R = b, \Delta = 0\right\} = \frac{1+c_b}{2+c_b} > 1/2$. Averaging over Δ then leads to $T < 0$. The example does not depend on the values of a_r and holds when $a_w = a_b$.

A special case arises when we impose the restrictions (a) $V \perp (\Delta, R)$ and (b) $E[\Delta|w,V] = E[\Delta|b,V]$, so that T in (A.35) becomes

$$T = P\left\{\psi(V) \leq g_{z,w}(V)\right\} - P\left\{\psi(V) \leq g_{z,b}(V)\right\} \ . \tag{A.39}$$

It follows immediately from this expression that if judge z is racially τ-unbiased, so $g_{z,w}(V) = g_{z,b}(V)$, then $T = 0$. It also follows from the same expression that if judge z is globally τ-biased against black defendants, so $g_{z,w}(V) > g_{z,b}(V)$, then $T > 0$. However, even in this very special case, one may still obtain $T = 0$ when $g_{z,w}(V) \neq g_{z,b}(V)$.