Reply to the Comment of Arnold, Dobbie, and Yang (2020)

Ivan A. Canay* Magne Mogstad† Jack Mountjoy‡

September 14, 2020

Abstract

This note addresses the "Comment on Canay, Mogstad, and Mountjoy (2020)" by Arnold, Dobbie, and Yang (ADY), which is appended to this reply. We divide the arguments into three points, all of which are puzzling. First, we do not mischaracterize the definition of racial bias in the published version of ADY. If the authors wrote the published definition, but actually meant a substantially different definition (such as the one that now appears in the new "Correction Appendix," also appended to this reply), then that is clearly the relevant mischaracterization. Second, focusing on clear-cut cases of (un)biased behavior is a feature of our argument, not a bug. The point is that even in the starkest, most unambiguous cases of unbiased and biased behavior, the outcome test can deliver the wrong conclusion. This logical invalidity of the outcome test also extends to intermediate cases where judges are biased against some defendants but not others. Third, to restore the logical validity of the outcome test, instead of invoking a decision model that justifies the test, ADY choose to redefine racial bias. Problematically, their substantial post-publication change in the definition of (un)biased judge behavior matters greatly for the interpretation and implications of their findings. The new definition is reverse-engineering, difficult to justify, and at odds not only with the work by Becker that ADY cite frequently, but also with more recent work by a subset of the authors of ADY.

*Department of Economics, Northwestern University. iacanay@northwestern.edu

†Department of Economics, University of Chicago; NBER. magne.mogstad@gmail.com

‡Booth School of Business, University of Chicago; NBER. jack.mountjoy@chicagobooth.edu

1 On the Definition of Racial Bias in ADY

The comment by ADY, appended to this reply, claims that we have mischaracterized their definition of bias, which appears in Definition 1, p. 1893, of the published paper (Arnold et al., 2018), hereafter ADY. The definition (using the notation in ADY) reads:

Definition 1 (ADY). We define judge j as racially biased against black defendants if $t_W^j(\mathbf{V}_i) > t_B^j(\mathbf{V}_i)$. Thus, for racially biased judges, there is a higher perceived benefit of releasing white defendants than releasing observably identical black defendants.

where "the perceived benefit of release for defendant i assigned to judge j is denoted by $t_r^j(\mathbf{V}_i)$, which is a function of observable case and defendant characteristics $\mathbf{V}_i$."

This definition has two features that are worth highlighting. First, the $\mathbf{V}_i$ that appears is identical on both sides of the inequality, and is indexed by the same i, which denotes a given defendant. A defendant i in the ADY setup is either black or white: defendant race $r \in W, B$ as written on p. 1893. Therefore, it seems natural to read the inequality in Definition 1 as a thought experiment: suppose we fixed defendant i's non-race characteristics $\mathbf{V}_i$ and switched defendant i's race from white to black. If $t_W^j(\mathbf{V}_i) > t_B^j(\mathbf{V}_i)$, then we call judge j racially biased. It is difficult to imagine other interpretations of this definition. The second sentence of Definition 1 even explicitly confirms that it involves white and black defendants who are otherwise "observably identical," i.e. share the same value of $\mathbf{V}_i$. It seems clear, then, that Definition 1 in ADY, as written in the published paper, is one where $\mathbf{V}_i$ is the *same on both sides* of the inequality.

Second, Definition 1 does not explicitly say anything about other defendants who are not this particular defendant i. But given that the definition always refers to plural "defendants," and the fact that the empirical test will pool all individuals in the sample, it would be strange to interpret i as anything besides generic: we could pick an arbitrary i and apply this definition. In other words, this definition holds for all i, which means it holds for all values in the empirical support of $\mathbf{V}$. Again, it is difficult to imagine other interpretations of this definition. If the authors did intend some other interpretation, like this definition only holding for some i, or some values of $\mathbf{V}$, or some other subset of defendants, but not for others, we would have expected the authors to clearly flag such caveats in the definition.

Therefore, we read the definition in the published paper carefully and interpreted it as we would expect any other reader to interpret it. If the authors actually meant to write a different definition (as stated in the new Correction Appendix, appended to this reply), then we will make sure to mention this correction, and its alternative definition, in future revisions of our paper. But the problem here is not us mischaracterizing a definition in a published paper that admits few, if any, alternative interpretations. The problem is that the text of the published paper features a substantially different definition than the one the authors supposedly intended. The responsibility for this clearly lies with the authors of the published paper, not with us as readers of it.

2 On the Definition of Racial Bias in CMM

In Canay et al. (2020), hereafter CMM, we use the following definition of racial bias (now moving to CMM notation):

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

In their comment, ADY make two main critiques of this definition. Both are actually irrelevant for our results. First, ADY are concerned that Definition 2.1 is "incomplete," in that it does not classify judges who are biased for some values of v and not biased for other values v. Indeed, our definition intentionally contains "if" statements, not "only if" or "if and only if" statements. This allows us to focus on the most clear-cut cases of unbiased and biased judges, without having to take a stand on situations where judges are prejudiced against some defendants but not others.

Theorem 3.1 in CMM shows why it is sufficient, and compelling, to focus on these clear cut cases to prove logical invalidity of the outcome test. Part (i) of Theorem 3.1 considers a judge at one end of the spectrum: racially unbiased for all values of v. We show that such a judge may nonetheless release marginal white defendants with higher misconduct rates than marginal black defendants, violating the logic of the outcome test. Part (ii) of Theorem 3.1 considers a judge at the other end of the spectrum: biased against black defendants for all values of v. We show that such a judge may nonetheless release marginal white defendants with equal or lower misconduct rates than marginal black defendants, again violating the logic of the outcome test.

The point is that even in the starkest, most clear-cut cases of unbiased and biased behavior, the outcome test can deliver the wrong conclusion. Further examination of more nuanced cases will not overturn this result. On the contrary, suppose that judge z satisfies

$$\tau(z,w,v) > \tau(z,b,v) \quad \text{for all } v \in \mathcal{V}_1 \quad \text{and} \quad \tau(z,w,v) = \tau(z,b,v) \quad \text{for all } v \in \mathcal{V}_2 \ , \qquad (1)$$

for some arbitrary partition of the support of V into $\mathcal{V}_1$ and $\mathcal{V}_2$. This judge is biased against black defendants for some values of v but not for others. The partition is arbitrary, so it can capture many different types of scenarios, including focusing on bias against defendants around the judge's margins of release. A version of Theorem 3.1 in CMM applies to this case too, and shows that such a judge "partially" biased against black defendants may nonetheless release marginal white defendants with equal or lower misconduct rates than marginal black defendants, again violating the logic of the outcome test. We will be sure to clarify this point in future revisions of our paper.

The second critique of Definition 2.1 is that "it rules out de-facto bias coming from seemingly non-race characteristics." To illustrate this point, ADY's comment offers several examples in which V is highly correlated with race, such that a judge who perceives different release benefits across different values of V is effectively perceiving different release benefits across race. This critique is also irrelevant for our results, since broadening the definition of racial bias to include judge preferences for any components of V that correlate with race does nothing to restore the logical

validity of the outcome test. We discuss this point explicitly on p. 17 of CMM: "while we do not necessarily object to this broadening of the definition of racial bias, it is important to observe that such broadening does not solve the problem at hand. Crucially, the presence of judge biases [or any non-misconduct considerations that vary] with respect to any non-race defendant characteristics can invalidate the outcome test for racial bias *even if those characteristics do not correlate with race or interact with race in judge decision making*."

To see this perhaps surprising result clearly, note that our framework (like ADY) does not impose any particular statistical relationship between defendant race R and non-race characteristics V. Our results therefore apply even if V contains only case and defendant characteristics that are completely independent of race. In this case, a judge z with a benefit function $\tau(z, v)$ is not only unbiased by Definition 2.1, since $\tau(\cdot)$ does not vary with r for any fixed v, but also sets release thresholds that are completely independent of defendant race, since V independent of R implies $\tau(z, V)$ is independent of R across defendants facing judge z. And yet, Theorem 3.1 in CMM still applies, so the outcome test may deem such a judge racially biased.

3 On the New Definition of Bias in ADY's Correction Appendix

To restore the logical validity of the outcome test, instead of invoking an extended Roy model of judge decision making, ADY choose to redefine racial bias. This new definition appears in a "Correction Appendix," appended to this reply. Redefining what it means for a judge to be biased is obviously a fundamental alteration to a paper titled "Racial Bias in Bail Decisions." A natural question is whether the authors' post-publication change in the definition of (un)biased judge behavior matters for the interpretation and implications of their findings. The answer is clearly yes, given our analysis in CMM.

To be precise, the new definition of ADY (in the notation of CMM) states that judge z is racially biased against black defendants if

$$\tau(z, w, V^*_{z,w}) > \tau(z, b, V^*_{z,b}) \ , \tag{2}$$

where $V^*_{z,w}$ denotes the non-race characteristics of judge z's marginal white defendant and $V^*_{z,b}$ denotes the non-race characteristics of judge z's marginal black defendant. As ADY acknowledge in their comment, $V^*_{z,w}$ and $V^*_{z,b}$ will generally *not* be the same. ADY's new definition of bias therefore involves comparisons of white and black defendants with different non-race characteristics. In contrast, the definition of bias in the published paper of ADY explicitly involved comparisons of white and black defendants with the same non-race characteristics.[1]

We are puzzled by this new definition for a few different reasons. First, at the outset, it is useful to recall that the work on taste-based discrimination by Becker (1957), which ADY extensively cite,

[1] As discussed above and in CMM, one can broaden the definition of racial bias to include preferences for non-race characteristics that correlate with race. Then the appropriate comparison would be between blacks and whites with the same values of the non-race characteristics that are independent of race.

shows that such discrimination may depress the wages (or employment) of black individuals relative to those of *equally productive* whites. Becker does not argue that one can draw conclusions about discrimination from differences in wages between blacks and whites of different productivity. In fact, throughout his analysis in *The Economics of Discrimination*, Becker reiterates his assumption that workers of different races are perfect substitutes in production. If black and white workers are instead "imperfect substitutes, they may receive different wage rates even in the absence of discrimination" (p. 17).

Second, ADY's new definition can define starkly unprejudiced behavior as biased and starkly prejudiced behavior as unbiased. For example, a judge with an expected benefit function invariant by race, satisfying

$$\tau(z, r, v) = \tau(z, v) \quad \text{for all } v \in \mathcal{V} \ , \tag{3}$$

nonetheless can also satisfy (2). This is illustrated in Figure 1a in CMM. So, the new definition would label such a judge as biased when race does not even enter the benefit function $\tau(\cdot)$. Furthermore, recall from the previous section that this can occur even if V is statistically independent of race, such that this judge's perceived release benefits are completely independent of defendant race. A definition that deems such behavior biased seems problematic and counterproductive.

In the other direction, a judge may satisfy

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

and yet according to (2) could be labeled as unbiased or even biased in the opposite direction. We illustrate this in Figures 1b and A.3a in CMM. The comment of ADY argues that (4) is a rather stringent definition of racial bias: such a judge perceives higher release benefits for white defendants over black defendants who have the same non-race characteristics v, and this preference holds across all values of v. We would therefore not have expected ADY's preferred definition to label such behavior as unbiased or biased in the opposite direction.

Third, this new definition appears to be simple reverse-engineering. In CMM, we show that differences in the misconduct rates of marginal white and black defendants need not be informative about bail judge racial bias, using the definition of bias and generalized Roy decision model in the published version of ADY. In response, ADY have now generated a new definition of bias in which differences in the misconduct rates of marginal white and black defendants identify racial bias *by construction*, turning the logical validity of the outcome test into a tautology.

Finally, ADY's comment and Correction Appendix fail to mention that the decision model and definition of bias in the published version of ADY also appear in a newer, related paper on consumer lending (Dobbie et al., 2020), accepted at *The Review of Economic Studies*. The latter paper refers to ADY (2018) as featuring "standard models of bias from the previous literature," and writes down nearly identical definitions of bias in the Appendix, including the phrase "observably identical" to clarify the comparisons involved in the definition of bias. It is puzzling that such an unintended definition can both be considered "standard," and survive explicit reproduction, in

subsequent peer-reviewed work by an overlapping set of authors.

References

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

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

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

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

Comment on Canay, Mogstad, and Mountjoy (2020)

David Arnold* Will Dobbie† Crystal S. Yang‡

September 2020

In Arnold, Dobbie, and Yang (2018, ADY), we find that marginally released white defendants have higher rates of pre-trial misconduct than marginally released black defendants. We interpret these findings as evidence of racial bias against black defendants through the lens of the marginal outcome test originally developed by Becker (1957). Canay, Mogstad, and Mountjoy (2020, CMM) question the interpretation of our empirical findings and the logical validity of the marginal outcome test. However, CMM's conclusions are based on an incomplete definition of racial bias that is different from the one used in ADY. Under ADY's definition of bias, the marginal outcome test is logically valid and a useful tool for studying discrimination in real-world settings.

*UC San Diego. Email: daarnold@ucsd.edu
†Harvard Kennedy School and NBER. Email: will_dobbie@hks.harvard.edu
‡Harvard Law School and NBER. Email: cyang@law.harvard.edu

Arnold, Dobbie, and Yang (2018, ADY hereafter) find that marginally released white defendants have higher rates of pre-trial misconduct than marginally released black defendants, where the marginally released defendant can be understood as the last defendant that a judge is willing to release for whom the judge is indifferent between release versus detention. We interpret these findings as evidence of racial bias against black defendants through the lens of the marginal outcome test originally developed by Becker (1957).

In a recent working paper, Canay, Mogstad, and Mountjoy (2020, CMM hereafter) critique the marginal outcome test for racial bias used in ADY. In this note, we respond to the main thrust of CMM's comments, which is that the marginal outcome test is logically invalid without further restrictions because it might find differences in outcomes at the margin when a judge acts on accurate predictions but does not have different preferences across defendant race. Based on these claims, CMM state that their "results call into question [ADY's] conclusions about racial bias among bail judges" (CMM, abstract).

In this response, we explain that CMM's conclusions are based on a problematic and incomplete definition of bias that is different from the one used in ADY. CMM are only willing to label a judge as racially biased if the judge treats white and black defendants differently across the entire characteristic space. This means that a judge is not labeled as racially biased by CMM even if she treats only a small subset of defendants equally and is racially biased for the vast majority of defendants. CMM's definition of racial bias also rules out instances of illegal discrimination coming from non-race characteristics. CMM, therefore, would incorrectly label a judge as racially unbiased even if the judge acts with discriminatory animus through non-race characteristics such as neighborhood. By comparison, ADY use a definition of racial bias at the margin that is likely to yield the correct conclusion of racial bias in these examples under reasonable assumptions.

We begin by summarizing the marginal outcome test and what it tells us. We focus on a simplified version of the marginal outcome test that builds on Becker (1957), noting that several other models also deliver the marginal outcome test. Following the notation in ADY, let i denote a defendant and $\mathbf{V}_i$ denote all case and defendant characteristics considered by the bail judge, excluding defendant race r_i. The expected cost of release for defendant i conditional on non-race characteristics $\mathbf{V}_i$ and race r_i is equal to the expected probability of pre-trial misconduct $\mathbb{E}[\alpha_i|\mathbf{V}_i, r_i]$.

The perceived benefit of release for defendant i assigned to judge j is denoted by $t_r^j(\mathbf{V}_i)$, which is a function of non-race case and defendant characteristics $\mathbf{V}_i$. The perceived benefit of release $t_r^j(\mathbf{V}_i)$ may vary by race $r \in W, B$ to allow for judge preferences to differ for white and black defendants following taste-based models of discrimination such as Becker (1957).

Suppose release decisions are consistent with a decision rule where judge j will release defendant i if and only if the expected cost of pre-trial release is less than or equal to the perceived benefit of release:

$$\mathbb{E}[\alpha_i|\mathbf{V}_i, r_i = r] \leq t_r^j(\mathbf{V}_i) \tag{1}$$

Given this decision rule, defendant i of race r is marginal for judge j if the expected cost of release is exactly equal to the perceived benefit of release, i.e. $\mathbb{E}[\alpha_i^j|\mathbf{V}_i, r_i = r] = t_r^j(\mathbf{V}_i)$. Let the non-race characteristics of the marginal defendant for judge j and race r be denoted $\mathbf{V}_{i,r}^*$.

We simplify our notation moving forward by letting the expected cost of release for the marginal defendant be denoted by $\alpha_r^j = E[\alpha_i^j|\mathbf{V}_i = \mathbf{V}_{i,r}^*, r_i = r]$. We correspondingly define $t_r^{j*} = t_r^j(\mathbf{V}_{i,r}^*)$.

The marginal outcome test is then given by:

$$D_j = \alpha_W^j - \alpha_B^j \tag{2}$$

or the expected difference in pre-trial misconduct rates among marginal white and marginal black individuals.

It is straightforward to show that a finding of $D_j \neq 0$ is inconsistent with accurate statistical discrimination and race-neutral thresholds at the margin. This is because by definition:

$$\alpha_W^j > \alpha_B^j \iff t_W^{j*} > t_B^{j*} \tag{3}$$

so that a finding of $D_j > 0$ implies that judge j has a higher perceived benefit of releasing white defendants than black defendants at the margin, or under an alternative model, implies that she overestimates the cost of release for black defendants relative to white defendants at the margin. In ADY, we define judge j as racially biased against black defendants if $t_W^{j*} > t_B^{j*}$.

CMM's main argument is that the marginal outcome test is logically invalid without further restrictions because it might find differences in outcomes at the margin when a judge acts on accurate predictions but does not, in fact, have different preferences across defendant race. Below, we provide two main critiques of these findings. First, we show that CMM's definition of racial bias is different from the ADY definition of racial bias and cannot speak to the validity of the ADY outcome test or ADY's findings. Second, we show that CMM's definition of racial bias is incomplete in that it is unable to identify important instances of racial discrimination, including instances prohibited by U.S. law.

Comment 1: The CMM Definition of Bias is Different from the ADY Definition of Bias

CMM's conclusions are based on a different definition of bias and cannot speak to the validity of the ADY outcome test or ADY's findings. The outcome test is logically valid under ADY's definition of racial bias, as shown in the above section.

In the published version of ADY, we say: "We define judge j as racially biased against black defendants if $t_W^j(\mathbf{V}_i) > t_B^j(\mathbf{V}_i)$" (ADY, p. 1893). We have subsequently clarified in a correction appendix that we define judge j as racially biased against black defendants if $t_W^{j*} > t_B^{j*}$, where $t_r^{j*} = t_r^j(\mathbf{V}_{i,r}^*)$.

By comparison, the definition of racial bias that CMM attribute to ADY (Definition 2.1) is "We say judge z is racially unbiased if $\tau(z,r,v) = \tau(z,v)$ for all $v \in \mathcal{V}$. If $\tau(z,w,v) > \tau(z,b,v)$ for all $v \in \mathcal{V}$, we say judge z is racially biased against black defendants" (CMM, p. 8). In CMM, $\tau(z,r,v)$ is the expected benefit of release by judge z for a defendant of race r and characteristics v, which corresponds to ADY's $t_r^j(\mathbf{V}_i)$, defined above as the perceived benefit of release by judge j for defendant i of race r and characteristics $\mathbf{V}_i$.

From their definition of racial bias, CMM define an outcome test as logically valid (Definition 3.1) if: "We say that the outcome test is logically valid if and only if $\text{sign}(\Lambda(w, V^*_{z,w}) - \Lambda(b, V^*_{z,b})) = \text{sign}(\tau(z,w,v) - \tau(z,b,v))$ for all $v \in \mathcal{V}$ and $z \in \mathcal{Z}$" (CMM, p. 10). In CMM, $\Lambda(r,v)$ represents the expected cost of release for a defendant of race r and characteristics v, and defendants of race r with non-race characteristics equal to $V^*_{z,r}$ are marginal for judge z. In CMM, $\Lambda(w, V^*_{z,w}) - \Lambda(b, V^*_{z,b})$ corresponds to the marginal outcome test defined above in Equation (2).

Motivated by CMM, our correction appendix clarifies that ADY's definition of racial bias is at the margin, not for all $v \in \mathcal{V}$, which we see as a substantially different definition of bias than the one used in CMM. Our correction appendix also clarifies that ADY's definition of racial bias at the margin does not require that non-race characteristics be identical for white and black defendants at the margin.

The distinction between ADY's definition of bias at the margin and CMM's definition of racial bias is also clear in the context of the published paper. We indicate throughout the paper that our definition of racial bias is at the margin (i.e., not for all $v \in \mathcal{V}$). For example, in the introduction of ADY, we state that "racial animus leads judges to discriminate against black defendants *at the margin of release*" (emphasis added) (ADY, p. 1889). This definition of bias at the margin is repeated in several parts of the paper (ADY, pp. 1889, 1922, 1929).

In various parts of ADY, we also indicate that our definition of bias does not require holding fixed non-race characteristics $\mathbf{V}_i$ at the margin (i.e., $\mathbf{V}^*_{i,W}$ and $\mathbf{V}^*_{i,B}$ may be different). For example, in discussing how variation in non-race characteristics of black and white defendants may affect understandings of racial bias, we explain "Another extension to our model concerns two distinct views about what constitutes racial bias. The first is that racial bias includes not only any bias due to phenotype, but also bias due to seemingly nonrace factors that are correlated with, if not driven by, race. For example, judges could be biased against defendants charged with drug offenses because blacks are more likely to be charged with these types of crimes. Our preferred estimates are consistent with this broader view of racial bias, measuring the disparate treatment of black and white defendants at the margin for all reasons unrelated to true risk of pre-trial misconduct, including reasons related to seemingly nonrace characteristics such as crime type" (ADY, p. 1904). This idea is again repeated throughout the paper (ADY, pp. 1888, 1904-1905, 1929).

Comment 2: A Critique of CMM's Definition of Racial Bias

CMM's definition of racial bias (Definition 2.1) is incomplete because it is unable to identify important instances of racial discrimination, including instances prohibited by U.S. law.

The first problem is that CMM's definition of racial bias is unable to say anything about judges who are biased for some v and not biased for some other v'. According to CMM's definition of racial bias, judge z is racially unbiased if $\tau(z, r, v) = \tau(z, v)$ for all $v \in \mathcal{V}$ and judge z is racially biased against black defendants if $\tau(z, w, v) > \tau(z, b, v)$ for all $v \in \mathcal{V}$ (CMM, p. 8). Therefore, even if a judge only treats a small subset of white and black defendants equally, CMM do not conclude that the judge is racially biased. The incomplete nature of CMM's definition of racial bias means that it risks saying nothing about real-world decision-makers who are unlikely to fall neatly into these extreme definitions of biased and unbiased behavior.

For example, under CMM's definition of bias, a judge is not labeled as racially biased even if she is racially biased against all defendants for whom she has the discretion to release. In many jurisdictions, defendants charged with capital offenses (such as first-degree murder) are not entitled to pre-trial release. Thus, a judge may treat black and white defendants charged with first-degree murder equally but the judge may be racially biased against all other black defendants. Under CMM's definition of racial bias, this judge is not classified as racially biased. Thus, in ignoring these institutional details, CMM's definition of racial bias is limited in its usefulness to ADY's setting of bail decisions. By comparison, definitions of bias at the margin (such as the one used in ADY) are complete and would likely suggest such a judge is, in fact, racially biased against black defendants. This is also the correct conclusion under U.S. law as there is no requirement that an actor is racially biased for all $v \in \mathcal{V}$ to engage in illegal discrimination.

The same problem can emerge even if a judge has the discretion to release infra-marginal defendants. For example, suppose that a judge is biased against poor black defendants but not biased against rich black defendants, and that 99 percent of black defendants and white defendants are poor and 1 percent of black defendants and white defendants are rich, where the only v is whether a defendant is rich or poor. Under CMM's definition of racial bias, CMM would not be able to conclude that such a judge is racially biased. By comparison, definitions of bias at the margin (such as the one used in ADY) would likely suggest such a judge is, in fact, biased if both marginal white and black defendants are poor. We view this as a more sensible interpretation of the judge's behavior, as the judge is racially biased for 99 percent of the population. Similarly, consider other characteristics such as gender, with prior work suggesting that judges may be racially biased against black men but not black women in sentencing decisions (Starr 2015). In such a scenario, CMM would again not be able to conclude that such a judge is racially biased, even though the vast majority of defendants in the criminal justice system are men. By comparison, definitions of bias at the margin (such as the one used in ADY) would likely suggest such a judge is, in fact, biased if both marginal white and black defendants are more likely to be men.

A second, related problem is that CMM's definition of bias is so narrow that it rules out de-facto bias coming from seemingly non-race characteristics. Racial bias only exists according to CMM if judges perceive higher benefits of release for white defendants than for black defendants who are identical in their non-race characteristics v. This again can be seen in CMM's definition of racial bias, which fixes v. We find this assumption troubling because it is at odds with legal and

structural definitions of racial bias. These issues are best illustrated through a series of examples, both hypothetical and drawn from the real world.

Consider, for example, redlining, generally defined as the illegal practice of denying a creditworthy applicant a loan for housing in a particular neighborhood on a discriminatory basis (such as based on the race or ethnicity of its residents). If CMM were to include neighborhood in v, they may erroneously conclude no racial bias even if it exists. In the criminal justice context, suppose that police are more likely to stop individuals in a particular neighborhood precisely because they know that the neighborhood has a high concentration of black residents. If CMM were to include neighborhood in v, they may wrongly conclude that the police are not racially biased. But that conclusion is problematic and at odds with anti-discrimination law. Under the Equal Protection Clause in federal courts, a facially neutral policy that has a disparate racial impact and was motivated by discriminatory animus or intent is illegal discrimination.[1]

The same type of problem can emerge when we condition on neighborhood in bail decisions. Consider a case where there are two zip codes, where zip code adds no information about a defendant's risk of pre-trial misconduct. Suppose that 99 percent of the defendants from one zip code are black, while only 1 percent of the defendants from the other zip code are black. Suppose, then, that the judge sets a stricter standard of release for all defendants in the predominantly black zip code compared to the predominantly white zip code, precisely because of discriminatory animus. In this scenario, we (and the law) believe the judge is acting with racial bias. According to the definition of racial bias in CMM, however, this judge would be labeled as unbiased, as $\tau(z, w, v) = \tau(z, b, v)$ for all $v \in \mathcal{V}$ where the only v here is neighborhood. By comparison, definitions of bias at the margin (such as the one used in ADY) are likely to yield the correct conclusion of racial bias under reasonable assumptions.

In ADY, we examine bias at the margin and do not fix v because of these reasons. Judges could be biased against defendants charged with drug offenses because black individuals are more likely to be charged with these types of crimes, or biased against defendants from certain neighborhoods because black individuals are more likely to reside there (ADY, p. 1904).[2] While we understand CMM's stated goal is to identify "whether, and to what extent, these group-level disparities are driven by relevant differences in underlying individual characteristics, or by biased decision makers" (CMM, abstract), their chosen definition of bias is so narrow that it rules out many plausible forms of racial bias. At a minimum, CMM's definition requires more conceptualization and justification. More broadly, we believe that economists must critically examine the notion that there must exist "relevant differences" across groups that can "explain" away observed racial differences when studying bias and discrimination.[3]

[1] See, e.g., *Hunter v. Underwood*, 471 U.S. 222, 233 (1985) (holding a facially race-neutral Alabama law disenfranchising those convicted of certain crimes invalid because it was enacted with a racially discriminatory purpose and had a racially disparate impact); see also *Arlington Heights v. Metropolitan Housing Development Corp.*, 429 U.S. 252 (1977).

[2] A detailed discussion of what it means to estimate the "effect of race" can be found in Sen and Wasow (2016).

[3] A thoughtful discussion of similar issues can be found in Professor William Spriggs' letter, available at

Summary: CMM claim that the marginal outcome test is logically invalid. However, CMM's conclusions are based on a different definition of racial bias than the one used in ADY. CMM's definition of racial bias is also incomplete in that it is unable to identify important instances of racial discrimination, including instances prohibited by U.S. law. Under ADY's definition of bias, the marginal outcome test is logically valid and a useful tool for studying discrimination in real-world settings.

References

[1] Arnold, David, Will Dobbie, and Crystal Yang. 2018. "Racial Bias in Bail Decisions." *Quarterly Journal of Economics*, 133(4): 1885–1932.

[2] Becker, Gary S. 1957. *The Economics of Discrimination*. University of Chicago Press.

[3] Canay, Ivan, Magne Mogstad, and Jack Mountjoy. 2020. "On the Use of Outcome Tests for Detecting Bias in Decision Making." Becker Friedman Institute Working Paper No. 2020-125.

[4] Sen, Maya, and Omar Wasow. 2016. "Race as a Bundle of Sticks: Designs that Estimate Effects of Seemingly Immutable Characteristics." *Annual Review of Political Science*, 19: 499–522.

[5] Starr, Sonja. 2015. "Estimating Gender Disparities in Federal Criminal Cases." *American Law and Economics Review*, 17(1): 127–159.

`https://www.minneapolisfed.org/~/media/assets/people/william-spriggs/spriggs-letter_0609_b.pdf?la=en`.

Correction Appendix to "Racial Bias in Bail Decisions"

David Arnold* Will Dobbie† Crystal S. Yang‡

September 2020

This document makes precise the formal definition of racial bias in our article "Racial Bias in Bail Decisions" published in the *Quarterly Journal of Economics* in November 2018. Our paper defines a judge as racially biased if they perceive a higher threshold of release for black defendants than white defendants at the margin, or under an alternative model, if they overestimate the cost of release for black defendants relative to white defendants at the margin. We refer to this verbal definition repeatedly throughout the paper. However, our formal definition of bias was insufficiently precise as to our intended definition of bias, which we realized in light of a recent working paper by Canay, Mogstad, and Mountjoy (2020).

To make our intended definition of bias clear, we make the following amendments to the published paper, where page numbers refer to the published version:

(1) On p. 1893, at the end of the paragraph beginning "The perceived benefit of release for defendant i..." we add the following definitions: "Let the non-race characteristics of the marginal defendant for judge j and race r be denoted $\mathbf{V}^*_{i,r}$. We correspondingly define $t^{j*}_r = t^j_r(\mathbf{V}^*_{i,r})$."

(2) On p. 1893, Definition 1 should be: "DEFINITION 1. Following Becker (1957, 1993), we define judge j as racially biased against black defendants if $t^{j*}_W > t^{j*}_B$. Thus, for racially biased judges, there is a higher perceived benefit of releasing white defendants than black defendants at the margin."

(3) On p. 1894, at the end of the sentence beginning "Given this decision rule...," $t^j_r(\mathbf{V}_i)$ should be $t^j_r(\mathbf{V}^*_{i,r})$ and at the end of the sentence beginning "We simply our notation...," α^j_r should be $\alpha^j_r = \mathbb{E}[\alpha^j_i | \mathbf{V}_i = \mathbf{V}^*_{i,r}, r_i = r]$.

(4) On p. 1895, Definition 2 should be: "DEFINITION 2. We define judge j as making racially biased prediction errors in risk against black defendants if $\tau^j_W(\mathbf{V}_i = \mathbf{V}^*_{i,W}) > \tau^j_B(\mathbf{V}_i = \mathbf{V}^*_{i,B})$. Thus, judges making racially biased prediction errors systematically overestimate the true cost of release for black defendants relative to white defendants at the margin."

(5) In Equations (4), (5), (6), (8) and any discussion of these equations, t^j_r should be t^{j*}_r.

*UC San Diego. Email: daarnold@ucsd.edu
†Harvard Kennedy School and NBER. Email: will_dobbie@hks.harvard.edu
‡Harvard Law School and NBER. Email: cyang@law.harvard.edu

(6) On p. 1922, in the sentence beginning with "Bail judges could, for example, harbor..." the phrase "observably similar" should be struck.