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). The principal contribution of ADY, relative to the past literature, was to develop new empirical methods to identify the outcomes of marginally released defendants as required by the marginal outcome test.

In a recent working paper, Canay, Mogstad, and Mountjoy (2020, CMM hereafter) critique the marginal outcome test for racial bias used in ADY. We are grateful to CMM for pushing the scientific frontier of this literature, as much work remains to be done in combating pervasive discrimination in America and elsewhere. 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 $t^j_r(\mathbf{V}_i)$, which is a function of non-race case and defendant characteristics $\mathbf{V}_i$. The perceived benefit of release $t^j_r(\mathbf{V}_i)$ 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 a judge as racially biased if her decisions cannot be solely explained by accurate statistical discrimination. Formally, judge j is 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 our online 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 online 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 online appendix also clarifies that ADY's definition of racial bias 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 and CMM's definition of racial bias is also clear in the context of the published paper. First, we indicate throughout the paper that we define a judge as racially biased when her decisions cannot be solely explained by accurate statistical discrimination, which can be identified by a finding of $D_j \neq 0$, as defined in Equation (2) above. In the introduction of ADY, for example, we state "In our setting, the outcome test is based on the idea that rates of pretrial misconduct will be identical for marginal white and marginal black defendants if bail judges are racially unbiased and the disparities in bail setting are solely due to accurate statistical discrimination. In contrast, marginal white defendants will have higher rates of pretrial misconduct than marginal black defendants if these bail judges are racially biased against blacks, whether that racial bias is driven by racial animus, inaccurate racial stereotypes, or any other form of racial bias" (ADY, p. 1886). This definition is repeated in several parts of the paper (ADY, pp. 1888, 1896, 1929).

Second, 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).

Third, 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 $\tau(z, w, v) = \tau(z, b, v)$

[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).

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]

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." NBER Working Paper No. 27802.

[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.

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