OUTCOME TESTS FOR POLICIES

CHANGHWA LEE[†], MALLESH M. PAI[‡], AND RAKESH VOHRA[†]

DECEMBER 7, 2021

ABSTRACT: The marginal outcomes test (Becker (1957)) has become a 'go-to test' of (un-)fairness/disparate impact in classification or allocation settings. We consider settings with two key properties: (1) the underlying attribute of the agent being classified is strategically chosen by the agent, and (2) the adjudicator commits to a classification *policy*, taking into account strategizing by the agent. In this setting we show the outcome test is misspecified: the optimal rule will result in different marginal outcomes across demographics, even in the absence of any discriminatory motive for the principal. We derive a correctly specified test in such a setting. The test statistic requires estimation of both marginal and average outcomes—the latter portion captures the effect on agents' incentives.

KEYWORDS: marginal outcome tests, discrimination.

JEL CLASSIFICATION: D63, D82, K40

[†]DEPARTMENT OF ECONOMICS AND DEPARTMENT OF ELECTRICAL & SYSTEMS ENGINEERING, UNIVERSITY OF PENNSYLVANIA

[‡]DEPARTMENT OF ECONOMICS, RICE UNIVERSITY

Lee and Vohra gratefully acknowledge financial support from NSF grant CCF-1763307. Pai gratefully acknowledges financial support from NSF grant CCF-1763349.

1. INTRODUCTION

There is much interest in evaluating the "fairness" of various socioeconomic institutions, e.g. criminal justice, access to employment/credit/education etc. In practice, this often boils down to focusing on a specific binary decision,[1] and comparing if this differs across various demographics e.g. black vs white defendants, male vs female job applicants. Within the economics literature, and more generally, the "gold standard" is the *marginal outcome test*, originally due to Becker (1957). A failure of this test is interpreted as evidence of discrimination by the decision maker (see e.g. Hull (2021), Bohren, Haggag, Imas, and Pope (2019)). It has been applied to a wide variety of settings.[2]

In this paper, we revisit the question: when is the marginal outcome test valid? We identify a natural class of settings of interest where a "fair" principal would choose a rule that fails the marginal outcome test, and identify the correct test for such settings. Specifically, these are settings where the principal is choosing a *policy*, and agents are responding *strategically* to the chosen policy. Settings that satisfy the desiderata we describe are easily motivated in practice. The idea that agents' relevant choices may be strategic and may respond to policy choices of the decision maker is of course standard in economics, and has long been considered in related settings (e.g. the design of affirmative action policy, see e.g. Coate and Loury (1993), Foster and Vohra (1992) or Fryer Jr and Loury (2013)) but largely absent from the literature on evaluating fairness. Settings where the adjudicator is making a policy choice also abound. For example, in the case of traffic stops it may amount to guidance issued by the leadership directing troopers on whom to stop. Similarly, as decision-making gets increasingly automated by the use of computers/ machine learning/ AI, it may be the choice of algorithm by the institution (e.g. the use of automated rules to determine who gets issued a loan in a banking setting, or the use of resume scanning software by an employer to determine which applicants get called back for an interview).

To fix ideas, let us first outline the model that the marginal outcome test implicitly assumes, focusing on the example of checking for racial bias in traffic stops by state troopers for contraband. A set of motorists each has a payoff-relevant attribute that is not directly observed by the decision maker (whether or not they are carrying contraband). The decision maker observes information about the motorist (including their race) and makes a binary decision on whether to interdict. Once the decision is made, this attribute is observed (i.e. upon conducting a traffic stop, the trooper learns whether the motorist was

[1] For example a judge choosing whether to acquit/ convict a defendant, a bank choosing whether or not to extend a loan to a loan applicant, an employer deciding whether or not to employ a job candidate.

[2] Some notable examples include: in the context of lending (Ferguson and Peters, 1995), judicial decision making (Arnold, Dobbie, and Yang (2018), Alesina and La Ferrara (2014)), traffic stop/ search decisions (Knowles, Persico, and Todd, 2001; Anwar and Fang, 2006; Antonovics and Knight, 2009) etc.

carrying contraband). The null hypothesis of no discrimination is that conditioned on being *marginal*, i.e. conditioned on the information seen by the decision maker being such that they are indifferent, the distribution of outcomes should be similar across races—after all, ceteris paribus, a decision maker should be indifferent at roughly the same rate of successful interdiction. Differences are either the result of a preference by the decision maker to pull over e.g. black drivers at a higher rate ("taste-based discrimination") or of an incorrect statistical model that causes the decision maker to over-estimate the risk of (marginal) black drivers ("incorrect statistical discrimination"). The underlying economic logic of the test is clear and uncontroversial, and therefore seemingly universally applicable.[3]

Formally, we show that marginal outcome tests may fail when the outcome of the agent is not exogenously determined, but instead depends on a strategic choice made by the agent (e.g. in our running example, the agents choose whether or not to carry contraband). In particular, suppose the decision maker chooses and commits to a decision policy *a priori*, and the agent understands this policy at the time of their own choice. In the language of Game Theory, the decision maker is a Stackelberg leader, or, equivalently, in the language of mechanism design, the decision maker has commitment. The agent's choice is thus based on a cost-benefit calculation given decision maker's policy (e.g. both the benefits of carrying contraband, and the associated risk of being apprehended). Therefore, the decision maker announces a policy that optimizes an objective function, taking into account that agents will respond to the underlying policy.

Our main positive result shows how to test for discrimination in such settings. At a high level, the intuition for this test can be described thusly: Under mild assumptions (Assumption 2), we show that the principal's optimal policy remains a group specific threshold on the signal. We show that agents of the two groups who generate a signal exactly equal to this threshold (i.e., the analog of the marginal agent at the standard marginal outcomes test) nevertheless will have different distributions of outcomes. This is precisely because the choice of threshold by the principal also affects agents' incentives. Since the optimal policy of the principal accounts also for how it affects the choices of the agents, we can derive a novel test statistic that a "fair" principal would equate across the groups.

The remainder of the paper is organized as follows—Section 2 outlines the general model, identifying in Section 2.1 some examples of special interest. Section 3 presents our results and discusses some of the key assumptions. Section 4 concludes with a discussion of the related literature.

[3]Operationally, one still needs to (correctly) identify the marginal agent which can be difficult in practice. The marginal outcomes test may also fail in richer models, see e.g. Canay, Mogstad, and Mountjoy (2020) which we discuss below.

2. MODEL

There is a set of agents. For each, the principal must take a binary decision $d \in D = \{0, 1\}$. This decision is the object of study—it could be, for example, traffic stops of motorists, loan approval/denial decisions, or job interview callback/rejection decisions etc.

Each agent belongs to a group $g \in \mathcal{G}$. A group corresponds to an observable characteristic of the agent, for instance race or gender, with respect to which we wish to evaluate the fairness of the principal's decision. We will concern ourselves with two groups, i.e. $\mathcal{G} = \{1, 2\}$, the extension to more than two groups is obvious.

Unobserved by the principal is a binary action choice by the agent $a \in \mathcal{A} = \{0, 1\}$. This action affects both the principal and the agent. In the traffic stop example, a is the choice of the agent on whether or not to carry contraband. In the case of employment, a might represent the choice of an agent to invest in human capital.

Prior to making their decision the principal observes the group identity of the agent. The principal also observes other information about the agent. Instead of directly modeling the information observed by the principal, we summarize this as a signal $s \in S \subseteq \mathbb{R}$ which is informative of the agent's action. The distribution of the signal depends only on the agent's chosen action and possibly their group. In particular, the signal is distributed according to CDF F_g^a (with pdf f_g^a) for an agent of group g who has taken action $a \in \mathcal{A}$.[4] We assume that the signals are informative in the same direction across groups, formally:

ASSUMPTION 1. *We assume that the distributions $\{f_g^a\}_{a \in \mathcal{A}}$ satisfy the Monotone Likelihood Ratio Property (MLRP), i.e. $\frac{f_g^1(s)}{f_g^0(s)}$ is non-decreasing in s for all groups.*[5]

The principal has a utility function $u : D \times \mathcal{A} \to \mathbb{R}$. By assumption therefore, the principal's decision and the agent's action are payoff relevant to the principal. Other observables, such as the agent's group identity and the signal they generate are payoff irrelevant by assumption (though of course they are informationally relevant in choosing an appropriate decision).

The choice for the principal is a *decision rule*, i.e., what decision $d \in D$ they make as a function of what they observe $(g, s) \in \mathcal{G} \times S$. We denote the decision rule by $\beta_g : S \to D$, i.e. $\beta_g(s)$ denotes the decision on an agent of group g for whom signal s was observed.

As we presaged above, the action choice of the agent is endogenous, and depends on the decision rule chosen by the principal. To be precise, agents also have preferences over action and decision, $v : D \times \mathcal{A} \times \Theta \to \mathbb{R}$, where Θ are payoff relevant types. The

[4] Note that we implicitly assume that the distribution of signals admits a density. Distributions with atoms etc. can be accommodated at some notational cost.

[5] If the distribution of signals is the same across groups, then this assumption is vacuous—it can be achieved by e.g. renaming signals appropriately.

distribution of types Θ in group g is given by μ_g. An agent of group g with type θ, facing the principal's decision rule β_g, chooses the action that maximizes their expected utility, i.e.

$$a_g^*(\theta, \beta_g) = \operatorname*{argmax}_{a \in \mathcal{A}} \int_S v(\beta_g(s), a, \theta) f_g^a(s) \mathrm{d}s.$$

The principal's problem then is to solve for each group:

$$\max_{\beta_g : S \to D} \int_\Theta \left(\int_S u(\beta_g(s), a^*(\theta, \beta_g)) \; f_g^{a^*(\theta, \beta_g)}(s) \mathrm{d}s \right) \; \mathrm{d}\mu_g(\theta) \; . \tag{OPT-g}$$

It will be useful, at this stage, to be clear about timing and observability. First, a principal, announces and commits to $\beta_g : S \to D$ for each group $g \in \mathcal{G}$. Then, each agent of group g privately observes their type θ (drawn according to distribution μ_g). The agent then chooses a utility maximizing action $a_g^*(\theta, \beta_g)$. Finally, for each agent, the principal observes the agent's group identity g and signal s (which depends on their chosen action), and takes the corresponding action, $\beta_g(s)$.

It will be useful to put some mild restrictions on the preferences of the principal and the agent to add structure to the model.

ASSUMPTION 2. *We make the following assumptions on the preferences of the principal and agent:*

(1) *Agent prefers decision* 1*: For any agent of any group g, type θ and action a, $v(1, a, \theta) \geq v(0, a, \theta)$.*
(2) *Principal prefers action* 1*: Ceteris paribus, the principal would prefer that agents take action* 1*, i.e., for any decision d, $u(d, 1) \geq u(d, 0)$.*
(3) *Principal prefers to match action and decision: $u(1, 1) \geq u(0, 1)$ and $u(0, 0) \geq u(1, 0)$.*

These assumptions are weak and capture the applications of interest: part (1) simply says that, ceteris paribus, decision 1 is the desirable decision from the agent's perspective (e.g. getting a loan, getting admitted to school, getting a job, not being pulled over in a traffic stop, etc). Similarly, part (2) says that from the principal's perspective, inducing action 1 by the agents is desirable (e.g. investing in human capital, not carrying contraband etc.). Finally, Part (3) says that the principal would like to match action and decision as much as possible. For example, in the traffic stop application, if action 1 is the agent's choice to not carry contraband (and 0 denotes the choice to carry contraband), the assumption simply says that the for an agent carrying contraband, the principal would prefer to interdict, while for an agent not carrying contraband, the principal would prefer not to interdict.

As we detail in examples below, this still allows flexibility. For instance this model accommodates the principal preferring that the agent take a particular action (e.g. the

design of education policy to maximize human capital investment by groups as in Fryer Jr and Loury (2013)).

2.1. *Examples*

Before proceeding to our results, we list some concrete examples of our model.

EXAMPLE 1 (Fixed Actions)**.** Our model subsumes the special case where agent actions are non-strategic, or equivalently for the purposes of the principal, the agent takes the action before the principal chooses their decision rule.

This can be achieved by giving agents a dominant action as a function of their type (i.e. their preferences over actions are independent of how the principal decides among agents). Formally, suppose $\Theta = \mathbb{R}$, with,

$$v(d, 1, \theta) = \theta, \quad \text{and} \quad v(d, 0, \theta) = -\theta.$$

EXAMPLE 2 (Strategic Agents)**.** Of course, more pertinent for our model is the case where agent's actions are strategically chosen to maximize the agents' expected utility given the principal's decision rule. A specific example of this is where $\Theta \subseteq \mathbb{R}_+$, and a given $\theta = (\theta_1, \theta_2)$ consists of two elements, where θ_1 represents the strength of the agent's preference to have decision $d = 1$ taken (e.g. additional value of getting a job), and θ_2 her net disutility of taking action $a = 1$ (e.g. disutility of investing in human capital). An agent of type θ has preferences given by:

$$v(d, a, \theta) = \chi_{\{d=1\}}\theta_1 - \chi_{\{a=1\}}\theta_2.$$

EXAMPLE 3 (Consequentialist preferences)**.** A special case that is relevant for some applications is where the principal only has preferences over the agent's action when they take decision $d = 1$. By a (slight) abuse of terminology, we call these consequentialist preferences: for example an employer only cares about the agent's choice of human capital investment if they choose to employ them ($d = 1$) but are otherwise indifferent. Formally, consequentialist preferences are preferences of the form $u(0, 0) = u(0, 1)(= 0)$, while $u(1, 0) \neq u(1, 1)$.

EXAMPLE 4 (Paternalistic Preferences)**.** Another natural special case to consider is one where the principal purely cares about the action taken by the agent—the decision is purely instrumental to incentivize the agent to take the desired action. For example, continuing with the employment/ human-capital application, these preferences might reflect those of a benevolent social planner wishing to maximize the fraction of agents who choose to invest in human capital. Formally, paternalistic preferences are of the form $u(d, 1) = 1, u(d, 0) = 0$.

3. Results and Discussion

To begin our analysis, note that the assumption on preferences (Assumption 2) combined with the assumption that the distribution of signals satisfies MLRP (Assumption 1) simplifies the principal's problem into a single threshold.

Lemma 1. *Under Assumptions 1 and 2, for each group g, the solution β_g to (OPT-g) simply specifies a threshold s_g^* such that $\beta_g(s) = 1 \iff s \geq s_g^*$.*

Proof. To see this, fix a group g and a decision rule of the principal $\beta_g(\cdot)$. Observe that any decision rule induces an effective probability p_g^a that an agent who takes action a receives decision 1, and correspondingly probability $1 - p_g^a$ of receiving decision 0. Note further by observation that agent's incentives are determined purely by p_g^1, p_g^0— any two decision rules that induce the same p_g^1, p_g^0 induce the same actions by the agent.

Next, note that by Assumption 1, for any feasible probabilities (p_g^1, p_g^0) that can be delivered by some decision rule, there exists a threshold rule which induces probabilities $(p_g^{1\prime}, p_g^{0\prime})$ such that $p_g^{1\prime} \geq p_g^1$ and $p_g^{0\prime} \leq p_g^0$. By Assumption 2 part (1), weakly more types of the agent take action 1 under this threshold rule than the original decision rule. By Assumption 2 part (2) and (3), this threshold rule can only be better in terms of the principal's objective (OPT-g) than the original. ■

Since the type θ of the agent is payoff irrelevant to the principal, as a function of the principal's threshold s_g^*, we can summarize the distribution of the agent's actions by a single number $\pi_g(s_g) \in [0,1]$. Here, $\pi_g(s_g^*)$ is the fraction of agents in group g that take action 1 when the principal uses a decision rule with threshold s_g^*. For the rest of this paper, we will assume that $\pi_g(\cdot)$ is a differentiable function.

In light of these simplifications, we can write the principal's problem as,

$$\max_{s_g^*} \quad u(1,1)(1 - F_g^1(s_g^*))\pi_g(s_g^*) + u(1,0)(1 - F_g^0(s_g^*))(1 - \pi_g(s_g^*)) \qquad \text{(Simple-Opt-g)}$$
$$+ u(0,1)F_g^1(s_g^*)\pi_g(s_g^*) + u(0,0)F_g^0(s_g^*)(1 - \pi_g(s_g^*))$$

This gives us the following (well-known) result, which justifies the validity of marginal outcome tests in settings where the actions of agents are fixed/ exogenously given (e.g. Example 1):

Observation 1. *Suppose agent's actions are fixed, i.e. $\pi_g(s_g^*) = \pi_g$ constant. Then, taking first-order conditions, the optimal threshold for the principal must satisfy*

$$0 = -u(1,1)f_g^1(s_g^*)\pi_g - u(1,0)f_g^0(s_g^*)(1 - \pi_g)$$
$$+ u(0,1)f_g^1(s_g^*)\pi_g + u(0,0)f_g^0(s_g^*)(1 - \pi_g).$$

$$\implies \frac{f_g^1(s_g^*)\pi_g}{f_g^0(s_g^*)(1-\pi_g)} = \frac{u(0,0)-u(1,0)}{u(1,1)-u(0,1)}.$$

The latter equation is the foundation of the marginal outcome test—after all the left hand side is the ratio of agents revealed to be taking action 1 *to action* 0 *among marginal agents; i.e. those that generate signal* s_g^* *where the principal is different between either decision. The first order condition asserts that this quantity must be equal across groups, since the right hand side is a quantity that is independent of group identity.*

However, in the general setting, the principal must also account for how their choice of threshold s_g^* affects an agent's behavior. The optimal thresholds for the principal's problem (Simple-Opt-g) will not equate marginal outcomes in general. Formally,

THEOREM 1 (Failure of the Marginal Outcome Test)**.** *Let* $\{s_g^*\}_{g\in\mathcal{G}}$ *be the solution to the principal's problem* (Simple-Opt-g)*. If* $\pi_g'(s_g^*) \neq 0$*, then for any other group* g'*, we have:*

$$\frac{f_g^1(s_g^*)\pi_g}{f_g^0(s_g^*)(1-\pi_g)} \neq \frac{f_{g'}^1(s_{g'}^*)\pi_{g'}}{f_{g'}^0(s_{g'}^*)(1-\pi_{g'})}.$$

In words, our theorem says that the standard statistic that is compared for marginal outcome tests may be different for different groups when agents' choices are endogenous and the principal's test is designed taking into account agents' responses. This despite the maintained assumption (by fiat) that the principal's preferences are independent of group identity.

So, in terms of positive results, what can we say about testing for discrimination in such a setting? As a first result, note that (Simple-Opt-g) already gives us a straightforward necessary first-order condition.

THEOREM 2. *Under the maintained assumptions, the solution to the principal's problem* (Simple-Opt-g) *for each group* g *must be a threshold* s_g^* *such that*

$$\begin{aligned} 0 = {} & (u(0,1)-u(1,1))\left(f_g^1(s_g^*)\pi_g(s_g^*) + F_g^1(s_g^*)\pi_g'(s_g^*)\right) \\ & + (u(0,0)-u(1,0))\left(f_g^0(s_g^*)(1-\pi_g(s_g^*)) - F_g^0(s_g^*)\pi_g'(s_g^*)\right) \\ & + (u(1,1)-u(1,0))\,\pi_g'(s_g^*). \end{aligned} \tag{FOC}$$

PROOF OF THEOREMS 1, 2. Theorem 2 follows from the assumption that $\pi_g(\cdot)$ is a differentiable function so that (FOC) is a necessary condition of optimality for (Simple-Opt-g). Theorem 1 follows by observation of (FOC) ■

Note that we can rewrite (FOC) as:

$$0 = (u(0,1) - u(1,1)) \frac{\mathrm{d}F_g^1(s_g^*)\pi_g(s_g^*)}{\mathrm{d}s_g^*} + (u(0,0) - u(1,0)) \frac{\mathrm{d}F_g^0(s_g^*)(1-\pi_g(s_g^*))}{\mathrm{d}s_g^*} + (u(1,1) - u(1,0))\, \pi_g'(s_g^*). \tag{FOC2}$$

Observe that this already provides a testable restriction if the econometrician knows the stated utility function of the principal. Testing this across groups therefore requires the econometrician to estimate, for each group g, quantities $\frac{\mathrm{d}F_g^1(s_g^*)\pi_g(s_g^*)}{\mathrm{d}s_g^*}$, $\frac{\mathrm{d}F_g^0(s_g^*)(1-\pi_g(s_g^*))}{\mathrm{d}s_g^*}$ and $\pi_g'(s_g^*)$. We discuss the possibility of such estimation in what follows. However, before this, we derive some implications for special cases.

COROLLARY 1. *Suppose the principal has consequentialist preferences of the form described in Example 3, i.e. $u(0,\cdot) = 0$. Then, for a principal applying the optimal policy, the optimal threshold s_g^* for any group solves*

$$u(1,1)\frac{\mathrm{d}(1-F_g^1(s_g^*))\pi_g(s_g^*)}{\mathrm{d}s_g^*} - u(1,0)\left(\pi_g'(s_g^*) + \frac{\mathrm{d}F_g^0(s_g^*)(1-\pi_g(s_g^*))}{\mathrm{d}s_g^*}\right) = 0 \tag{1}$$

i.e., under the maintained assumption about the nature of the principal's preferences, the ratio of

$$\frac{\mathrm{d}(1-F_g^1(s_g^*))\pi_g(s_g^*)}{\mathrm{d}s_g^*} \quad \textit{and} \quad \pi_g'(s_g^*) + \frac{\mathrm{d}F_g^0(s_g^*)(1-\pi_g(s_g^*))}{\mathrm{d}s_g^*}$$

is equal across groups.

COROLLARY 2. *Suppose the principal has paternalistic preferences of the form described in Example 4, i.e $u(\cdot,0) = 0$. For a principal applying the optimal policy, the optimal threshold s_g^* for any group solves*

$$0 = (u(0,1) - u(1,1)) \frac{\mathrm{d}F_g^1(s_g^*)\pi_g(s_g^*)}{\mathrm{d}s_g^*} + u(1,1)\pi_g'(s_g^*). \tag{2}$$

i.e., under the maintained assumption of paternalistic preferences, the ratio of

$$\frac{\mathrm{d}F_g^1(s_g^*)\pi_g(s_g^*)}{\mathrm{d}s_g^*} \quad \textit{and} \quad \pi_g'(s_g^*)$$

is equal across groups.

3.1. *Discussion*

Estimation. Our corollaries provide analogs of the marginal outcomes test under the assumption of strategic agents and a "mechanism designer" principal. The possibility to execute such a test depends on the ability to estimate the relevant quantities, i.e. $\pi_g'(s_g^*)$,

$\frac{\mathrm{d}F_g^1(s_g^*)\pi_g(s_g^*)}{\mathrm{d}s_g^*}$, and $\frac{\mathrm{d}F_g^0(s_g^*)(1-\pi_g(s_g^*))}{\mathrm{d}s_g^*}$. It is worth discussing what these quantities correspond to in terms of the underlying model.

Let us start with the first: $\pi_g'(s_g^*)$. This is the derivative of the fraction of group g agents taking the action 1 with respect to the principal's threshold for that group s_g^*. As we will see this is the novel term that would need to be estimated (relative to a traditional marginal outcomes test). Estimating this would either require further modeling of the agents' incentives (i.e. a structural model of their choices), or, e.g., identifying variation (e.g. different principals who use slightly different thresholds) that can be exploited. Of course, any such estimation would be nontrivial, so we do not speculate further here.

Given an estimate of $\pi_g'(s_g^*)$, the second term $\frac{\mathrm{d}F_g^1(s_g^*)\pi_g(s_g^*)}{\mathrm{d}s_g^*}$ is easier to estimate. Note that by an application of the product rule, it can be written as

$$f_g^1(s_g^*)\pi_g(s_g^*) + F_g^1(s_g^*)\pi_g'(s_g^*).$$

Here, the first term, $f_g^1(s_g^*)\pi_g(s_g^*)$ corresponds to the fraction of the agents at the principal's threshold (s_g^*) who have taken action 1—this is exactly the numerator of the standard marginal outcomes test (recall Observation 1) and can be estimated similarly. The second term is the product of $F_g^1(s_g^*)$ (the false negative rate implied by the principal's threshold) and the previously estimated $\pi_g'(s_g^*)$. Analogously, the third term $\frac{\mathrm{d}F_g^0(s_g^*)(1-\pi_g(s_g^*))}{\mathrm{d}s_g^*}$, by an appeal to the product rule, can be written as

$$f_g^0(s_g^*)(1-\pi_g(s_g^*)) - F_g^0(s_g^*)\pi_g'(s_g^*).$$

Here again, the first term is the denominator of the standard marginal outcomes test, while the second is the product of the previously estimated $\pi_g'(s_g^*)$ and $1-$ the false positive rate implied by the principal's threshold.

In short, therefore, relative to the standard marginal outcome test, three new quantities need to be estimated for each group: the previously discussed $\pi_g'(s_g^*)$; and the False Positive and False Negative rates for the group . Note that the first of these has no analog in the setting considered by the standard marginal outcome test. The latter two, i.e. the False Positive/ False Negative Rate of the principal's decision rule are well understood quantities economically: interestingly however these are precisely the quantities that the marginal outcomes test eschewed. The reason for their inclusion is simple: when agents' actions are endogenous, their incentives depend on the entire distribution of the principal's decisions, not just the principal's decisions at the margin.

The Role of Commitment. It may be useful at this stage to clarify the role of the two modeling assumptions we made, i.e. (1) commitment to the classification rule by the principal and (2) a relevant action being taken by strategic agents *after learning the principal's decision rule*.

First, as discussed, (2) is critical—if agents' actions are exogenously fixed then the marginal outcome test is valid independent of (1) (recall Observation 1).

As we show in what follows, (1) is also critical, i.e. in the absence of commitment, again, the marginal outcome test is valid. To see this, observe that in the absence of commitment, the principal can only take the sequentially rational decision at the time of deciding (i.e. the action that maximizes their expected utility conditional on the observed signal). Formally, a principal who sees signal s, in the absence of commitment takes $d = 1$ over $d = 0$ if:

$$u(1,1)\pi_g f_g^c(s) + u(1,0)(1-\pi_g) f_g^o(s) > u(0,1)\pi_g f_g^c(s) + u(0,0)(1-\pi_g) f_g^o(s).$$

Here π_g is the fraction of agents in group g who take action 1, since this is determined at the time the principal takes their action. The principal is indifferent if

$$u(1,1)\pi_g(s_g) f_g^c(s) + u(1,0)(1-\pi_g(s_g)) f_g^o(s) = u(0,1)\pi_g(s_g) f_g^c(s) + u(0,0)(1-\pi_g(s_g)) f_g^o(s)$$

which is equivalent to

$$\frac{f_g^1(s_g^*)\pi_g}{f_g^0(s_g^*)(1-\pi_g)} = \frac{u(0,0)-u(1,0)}{u(1,1)-u(0,1)}.$$

Under the MLRP assumption (assumption 1), the principal follows a threshold rule of decision $d = 1$ for $s > s_g^*$ and $d = 0$ otherwise. By observation, this is the same as the case of exogenously fixed actions, and the marginal outcome test remains valid. As we describe below, this case of endogenous actions but without commitment was considered in Knowles, Persico, and Todd (2001) and Anwar and Fang (2006).

4. Related Literature

The original marginal outcome test is generally attributed to Becker (1957). More recently, Hull (2021) and Bohren, Haggag, Imas, and Pope (2019) revisit the marginal outcomes test and provide formal models in which the test is valid. On the flip side, Canay, Mogstad, and Mountjoy (2020) point out that there are natural models in which the marginal outcome test fails in both directions, i.e. differences in marginal outcome are possible despite a principal who by assumption has no discriminatory preferences; and vice versa. Critically, they allow an agent's observable characteristics to directly enter the principal's preferences. This may be reasonable in some settings, nevertheless we follow the majority of the literature in assuming that other observables are informative for the principal but do not directly affect their preferences. Our "negative result" (i.e., Theorem 1) therefore is for conceptually different reasons.

As we pointed out earlier, a major difficulty operationalizing the marginal outcome test is correctly identifying the marginal agent so as to do the appropriate comparison. Various approaches have been taken to get around this. Closest in spirit to our paper is

the paper of Knowles, Persico, and Todd (2001) on detecting racial bias in traffic stops (see also the extensions in the appendix of Anwar and Fang (2006))—they construct an equilibrium model in which both agents and police officers are strategic. In the taxonomy of our model, these papers consider a setting with out commitment to a policy, i.e., one where the police officers take a sequentially rational action given the information they observe rather than committing *a priori* to a policy. Operationally, in the equilibrium of their model, the marginal and average outcomes for agents are the same (since agents are observationally homogeneous to police officers beyond their race). This allows them to construct a test based on the (easy to observe) average outcomes. A majority of the papers however take a non-structural approach. In particular they use quasi-experimental approaches to identity the marginal agent, for example the random assignment of judges to cases— see e.g. Arnold, Dobbie, and Yang (2018), Feigenberg and Miller (2020), Grau, Vergara, et al. (2020).

More recently, there has been progress towards more robust tests: see e.g. Marx (2018) or Martin and Marx (2021). These papers construct tests based on necessary implications of unbiased decision making— i.e. passing the test does not necessarily imply unbiased decision, but failing the test is (strong) evidence of prejudice.

The idea of commitment to a policy, though not stated as such, also arises when thinking of the design of e.g., a machine learning algorithm to automatically classify agents. Computer scientists have grown increasingly concerned about whether and how even seemingly neutral algorithms can treat different demographic groups differently. This has resulted in literatures studying the incompatibility of various formal notions of fairness (see Chouldechova (2017), Kleinberg, Mullainathan, and Raghavan (2016)). A subsequent literature has proposed (or criticized) notions of fairness based on ethical/ normative grounds and discussed the possibility of algorithms that are fair with respect to such notions (see e.g. Dwork, Hardt, Pitassi, Reingold, and Zemel (2012); Hardt, Price, and Srebro (2016); Corbett-Davies and Goel (2018); Corbett-Davies, Pierson, Feller, Goel, and Huq (2017); Feller, Pierson, Corbett-Davies, and Goel (2016); Friedler, Scheidegger, and Venkatasubramanian (2016); Kearns, Neel, Roth, and Wu (2018); Hébert-Johnson, Kim, Reingold, and Rothblum (2018); Liu, Simchowitz, and Hardt (2019)). Perhaps the closest to the present paper is the paper of Jung, Kannan, Lee, Pai, Roth, and Vohra (2020) who study the design of optimal policy with respect to a specific objective function (in our terminology, a principal with paternalistic preferences, Example 4), and derive the optimal classification rule.

Finally, as we pointed out earlier, there has also been a literature in economic theory trying to understand the design of (e.g. affirmative action) policy taking into account differing incentives in differing groups to take a relevant action (e.g., invest in human capital)— see e.g. Loury et al. (1977), Coate and Loury (1993), Foster and Vohra (1992) or

Fryer Jr and Loury (2013). The broader literature is surveyed in Fang and Moro (2011). Even outside the context of fairness/ discrimination, several papers study the provision of incentives in hiring/ admission settings. For an example of the former see Hatfield, Kojima, and Kominers (2014) or Hatfield, Kojima, and Kominers (2018) who point out that in employment matching settings, workers need to get the ex-post marginal product of their labor to align their incentives to undertake the ex-ante efficient investment in human capital. In the latter setting, Frankel and Kartik (2019) consider a setting where applicants have both a underlying ability and an ability to "game" the signal observed by the decision maker. They show that a decision maker wishing to match on underlying ability may wish to commit to a policy that conditions less strongly on the observed signal so as to disincentivize gaming. Finally, the work of Frankel (2021) studies a setting where a principal must delegate to an agent of unknown bias, and has limited control over the agent. For example, relevant to the present context, this could be a city hiring traffic police officers of unknown bias. The paper shows that the principal hires and delegates using a rule such that *marginal* police officer hired conducts traffic stops which would satisfy the marginal outcomes test.

REFERENCES

ALESINA, A., AND E. LA FERRARA (2014): "A test of racial bias in capital sentencing," *American Economic Review*, 104(11), 3397–3433.

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(1), 163–177.

ANWAR, S., AND H. FANG (2006): "An alternative test of racial prejudice in motor vehicle searches: Theory and evidence," *American Economic Review*, 96(1), 127–151.

ARNOLD, D., W. DOBBIE, AND C. S. YANG (2018): "Racial bias in bail decisions," *The Quarterly Journal of Economics*, 133(4), 1885–1932.

BECKER, G. S. (1957): *The economics of discrimination*. University of Chicago press.

BOHREN, J. A., K. HAGGAG, A. IMAS, AND D. G. POPE (2019): "Inaccurate Statistical Discrimination: An Identification Problem," Working Paper 25935, National Bureau of Economic Research.

CANAY, I. A., M. MOGSTAD, AND J. MOUNTJOY (2020): "On the use of outcome tests for detecting bias in decision making," Discussion paper, National Bureau of Economic Research.

CHOULDECHOVA, A. (2017): "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *arXiv preprint arXiv:1703.00056*.

COATE, S., AND G. C. LOURY (1993): "Will affirmative-action policies eliminate negative stereotypes?," *The American Economic Review*, pp. 1220–1240.

CORBETT-DAVIES, S., AND S. GOEL (2018): "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv preprint arXiv:1808.00023*.

CORBETT-DAVIES, S., E. PIERSON, A. FELLER, S. GOEL, AND A. HUQ (2017): "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806. ACM.

DWORK, C., M. HARDT, T. PITASSI, O. REINGOLD, AND R. ZEMEL (2012): "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226. ACM.

FANG, H., AND A. MORO (2011): "Theories of statistical discrimination and affirmative action: A survey," in *Handbook of social economics*, vol. 1, pp. 133–200. Elsevier.

FEIGENBERG, B., AND C. MILLER (2020): "Racial disparities in motor vehicle searches cannot be justified by efficiency," Discussion paper, National Bureau of Economic Research.

FELLER, A., E. PIERSON, S. CORBETT-DAVIES, AND S. GOEL (2016): "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear," *The Washington Post*.

FERGUSON, M. F., AND S. R. PETERS (1995): "What constitutes evidence of discrimination in lending?," *The Journal of Finance*, 50(2), 739–748.

FOSTER, D. P., AND R. V. VOHRA (1992): "An economic argument for affirmative action," *Rationality and Society*, 4(2), 176–188.

FRANKEL, A. (2021): "Selecting Applicants," *Econometrica*, 89(2), 615–645.

FRANKEL, A., AND N. KARTIK (2019): "Improving information from manipulable data," *Journal of the European Economic Association*.

FRIEDLER, S. A., C. SCHEIDEGGER, AND S. VENKATASUBRAMANIAN (2016): "On the (im) possibility of fairness," *arXiv preprint arXiv:1609.07236*.

FRYER JR, R. G., AND G. C. LOURY (2013): "Valuing diversity," *Journal of political Economy*, 121(4), 747–774.

GRAU, N., D. VERGARA, ET AL. (2020): "A Simple Test for Prejudice in Decision Processes: The Prediction-Based Outcome Test," *Santiago*.

HARDT, M., E. PRICE, AND N. SREBRO (2016): "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, pp. 3315–3323.

HATFIELD, J. W., F. KOJIMA, AND S. D. KOMINERS (2014): "Investment incentives in labor market matching," *American Economic Review*, 104(5), 436–41.

——— (2018): "Strategy-proofness, investment efficiency, and marginal returns: An equivalence," *Becker Friedman Institute for Research in Economics Working Paper*.

HÉBERT-JOHNSON, Ú., M. KIM, O. REINGOLD, AND G. ROTHBLUM (2018): "Multicalibration: Calibration for the (computationally-identifiable) masses," in *International Conference on Machine Learning*, pp. 1944–1953.

HULL, P. (2021): "What Marginal Outcome Tests Can Tell Us About Racially Biased Decision-Making," Discussion paper, National Bureau of Economic Research.

JUNG, C., S. KANNAN, C. LEE, M. PAI, A. ROTH, AND R. VOHRA (2020): "Fair Prediction with Endogenous Behavior," in *Proceedings of the 21st ACM Conference on Economics and Computation*, EC '20, p. 677–678, New York, NY, USA. Association for Computing Machinery.

KEARNS, M., S. NEEL, A. ROTH, AND Z. S. WU (2018): "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness," in *International Conference on Machine Learning*, pp. 2569–2577.

KLEINBERG, J., S. MULLAINATHAN, AND M. RAGHAVAN (2016): "Inherent Trade-Offs in the Fair Determination of Risk Scores," *arXiv preprint arXiv:1609.05807*.

KNOWLES, J., N. PERSICO, AND P. TODD (2001): "Racial bias in motor vehicle searches: Theory and evidence," *Journal of Political Economy*, 109(1), 203–229.

LIU, L. T., M. SIMCHOWITZ, AND M. HARDT (2019): "The Implicit Fairness Criterion of Unconstrained Learning," in *International Conference on Machine Learning*, pp. 4051–4060.

LOURY, G., ET AL. (1977): "A dynamic theory of racial income differences," *Women, minorities, and employment discrimination*, 153, 86–153.

MARTIN, D., AND P. MARX (2021): "A Robust Test of Prejudice for Discrimination Experiments.","" .

MARX, P. (2018): "An absolute test of racial prejudice," *The Journal of Law, Economics, and Organization*.