Can Policies Regulating Information Acquisition Alleviate Statistical Discrimination?*

Koda Gursoy[†1]

[1]Princeton University

1 Introduction

Affirmative action has long been a hotly contested issue in the political sphere and in the economic literature. Empirical research on its use, for instance, in higher education, returns mixed results on its efficacy (Arcidiacono et al., 2011; Bleemer, 2022). Some affirmative action policies seek to address discrimination without simply offering a strict boost to the score of some individuals' applications. One such example is the Rooney Rule in the NFL, which requires that the league's teams offer in-person interviews to a certain number of "minority" candidates for various coaching positions. This sort of regime aims to target discrimination resulting from information frictions.

In this paper, I aim to examine whether such policies are effective by building on theoretical models of statistical discrimination. I allow employers to choose any information structure from which to receive signals about worker productivity, then characterize solutions to the employer's and worker's problems and define equilibrium accordingly. I then provide a simple necessary and sufficient condition for the existence of discriminatory equilibria. Next, I examine three affirmative action policies regulating employer signal acquisition. The first requires sufficient confidence in lack of qualification before rejection, and always fails to prevent discrimination. The second requires sufficient confidence in qualification before hiring to the desirable task, and can prevent discrimination under certain model parameters, but may also make discrimination strictly worse. The third requires equal desirable-task hiring probability for qualified workers across groups, and always succeeds in preventing discrimination.

These results demand caution from those designing affirmative action policies regulating information acquisition. While this version of affirmative action may succeed at alleviating statistical discrimination, it may also make the problem worse. Before implementing any such affirmative action regime, regulators must carefully consider how the changes in information structure induced by their policies will alter human capital acquisition, so as not to exacerbate the very issues they seek to address.

*Thanks to Adam Kapor for advising this project, as well as Faruk Gul and Xiaosheng Mu for their helpful comments.

[†]Corresponding author: koda.gursoy@princeton.edu

2 Literature Review

In models of statistical discrimination, disparities in outcomes between groups arise not due to animus but from rational agents' use of group-level statistics when making decisions under imperfect information about individuals. Arrow (1971/1995) argues that employer beliefs about group productivity can be self-confirming. If firms expect a group to invest in qualification at lower rates, they optimally offer that group worse opportunities. Workers from that group then optimally invest at low rates, which validates the original belief. Coate and Loury (1993) formalize an Arrovian model of statistical discrimination with noisy signals and an affirmative action regulating the ex-ante likelihood of getting hired to a desirable task. They find that their policy may either eliminate or sustain discrimination.

My paper alters Coate and Loury's (1993) model of statistical discrimination in hiring, preserving much of the setup, but allowing employers to choose any signal structure to assess worker qualification status.[1] The introduction of flexible information structures and imposition of this affirmative action regime depart from both the Coate-Loury model and later static equilibrium models of statistical discrimination which assume an exogenous signal distribution and impose an affirmative action policy on employers at the hiring stage of the game (Fang and Moro, 2011).

To allow this flexible information acquisition, I draw from the literature on rational inattention, particularly Matějka and McKay (2015). I define an information structure to be a set of conditional distributions of a signal given a worker's qualification status as Matějka and McKay do, allowing employers total freedom to choose these distributions. I then characterize the posterior beliefs the signals induce given the employer's prior belief on qualification rate. Information structures are subject to a entropy reduction cost function as is standard in the rational inattention literature (Matějka and McKay, 2015). I simplify the employer's optimization problem by appealing to Matějka and McKay's result that any optimal information structure will have at most one posterior corresponding to each action the decision maker can take (2015, p. 272).

3 Model

As in Coate and Loury (1993), we begin with a continuum of identical employers of measure 1, and a continuum of workers of measure $m >> 1$. Workers have an observable group membership $g \in \{B, W\}$. The primary payoff-relevant action of the employer is to assign each worker to one of two jobs, denoted as task zero and task one. All workers can effectively complete task zero, but only qualified workers can successfully complete task one. Workers receive payoff w if they are assigned to task one, so all workers prefer to be assigned to task one independent of qualification status.

[1] After completing the original version of this paper, I became aware of Echenique and Li (2025), who independently study statistical discrimination under rational inattention. They model a tournament between two agents for a single promotion, whereas I preserve the Coate-Loury labor market structure. They focus on which equilibrium is most preferable to each agent and employ an affirmative action policy regulating likelihood of promotion; my work studies multiple affirmative action policies regulating the information acquisition stage explicitly.

Workers make the decision whether or not to become qualified, and the cost of becoming qualified is independently and identically distributed across all workers according to a log-concave distribution characterized by smooth CDF $F(c)$ which is equal to 0 at $c = 0$ and strictly increasing on $(0, w)$. We say a worker has qualification status $\omega \in \Omega \equiv \{Q, U\}$.

The employer gets payoff x_q $(-x_u)$ for assigning a qualified (unqualified) worker to task one. Employer and worker payoffs from assignment to task zero are normalized to zero. Thus, employers wish to assign a worker to task one if and only if they are qualified.

Employers hold a prior belief $\mu_G \in [0, 1]$, $G \in \{B, W\}$ about the proportion of each group that is qualified. Before workers are sent to employers to be assigned jobs, the employers must decide on an information structure from which to receive a signal about the qualification status of each worker. Employers select information structures independently for each group, and they optimize their selection of information structure based on their prior beliefs.

We allow employers to receive signals $s \in S \equiv [0, 1]$. An information structure is defined by a conditional distribution of s given qualification status ω for each $\omega \in \Omega$. The employer has total freedom to select these conditional distributions.

Definition 1 (Information Structures)**.** The set of information structures Π is the collection of all mappings $\pi : \Omega \to \Delta(S)$. We write the density of the conditional distribution of s given ω as $\pi(s|\omega)$.

It follows that for a prior belief μ, the unconditional distribution of s under μ is

$$\pi_\mu(s) = \mu \cdot \pi(s|Q) + (1 - \mu) \cdot \pi(s|U). \tag{1}$$

We then define the function γ mapping the signal received to the posterior probability of a worker being qualified, $\gamma_\mu(\pi) : S \to [0, 1]$, as

$$\gamma_\mu(s|\pi) \equiv \Pr(Q|s, \mu, \pi).$$

By Bayes' Law, we can precisely specify this posterior probability:

$$\gamma_\mu(s|\pi) = \frac{\mu \cdot \pi(s|Q)}{\mu \cdot \pi(s|Q) + (1 - \mu) \cdot \pi(s|U)}. \tag{2}$$

Information structures are priced by the entropy-based cost function standard in the rational inattention literature (Matějka and McKay, 2015):

$$K(\pi) \equiv \lambda(H(\mu) - E_s[H(\gamma_\mu(s|\pi)]), \tag{3}$$

where H is the Shannon entropy function:

$$H(\mu) = \begin{cases} -[\mu \cdot \log(\mu) + (1 - \mu) \cdot \log(1 - \mu)] & \mu \in (0, 1) \\ 0 & \mu = 0, 1 \end{cases}.$$

3.1 Unconstrained Employer's Problem

Without any affirmative action intervention, employers are free to select a different information structure and mapping rule from signals to task assignment for each group. Therefore, I will solve the employer's problem as though they are only selecting a strategy for one group given their prior belief on the qualification rate in that group. Since my definitions of an information structure and its cost function are the same as those in Matějka and McKay (2015), I leverage their results to simplify the employer's optimization problem.

Employers can take action $i \in A \equiv \{\text{Task Zero, Task One}\}$. Let a strategy of the employer be noted as (π, a), where $\pi \in \Pi$ is an information structure and $a : S \to A$ is an action rule mapping signals to task assignment.

Then by Lemma 1 in Matějka and McKay (2015), under any optimal strategy (π^*, a^*), if an action $i \in \{\text{Task Zero, Task One}\}$ is taken with positive unconditional probability, there exists a posterior $\gamma_i \in [0, 1]$ such that conditional on taking action i, the probability that the posterior belief is γ_i is 1.

Therefore we restrict our signal space to only two signals, $s \in \{0, 1\}$. Then, the employer needs only to choose $p_q = \pi(1 \mid Q)$ and $p_u = \pi(1 \mid U)$. The employer will assign all those who return a high signal, $s = 1$, to task one, and all those who return a low signal, $s = 0$, to task zero. That is, $a^*(0) =$ Task Zero, $a^*(1) =$ Task One. This fully captures the space of possible optimal strategies. $p_q = p_u = p$ represents a strategy in which the employer acquires no information and assigns share p of workers to task one, and $p_q \neq p_u$ is a strategy in which information is purchased. Given this, we can express the posterior beliefs corresponding to high and low signals in terms of p_q and p_u:

$$\begin{aligned} \gamma_h \equiv \gamma_\mu(1|p_q, p_u) &= \begin{cases} \mu & p_q = p_u \\ \frac{\mu p_q}{\mu p_q + (1-\mu)p_u} & p_q \neq p_u \end{cases}, \\ \gamma_l \equiv \gamma_\mu(0|p_q, p_u) &= \begin{cases} \mu & p_q = p_u \\ \frac{\mu(1-p_q)}{\mu(1-p_q) + (1-\mu)(1-p_u)} & p_q \neq p_u \end{cases}. \end{aligned} \tag{4}$$

Then, we can write the employer's objective function in terms of p_q and p_u.

$$R(p_q, p_u) = x_q \mu p_q - x_u(1-\mu)p_u - K(p_q, p_u) \tag{5}$$

for $(p_q, p_u) \in [0, 1]^2$, with cost function

$$K(p_q, p_u) = \lambda \left[H(\mu) - (\mu p_q + (1-\mu)p_u) H(\gamma_h) - (\mu(1-p_q) + (1-\mu)(1-p_u)) H(\gamma_l) \right]. \tag{6}$$

So, the employer's problem is to maximize $R(p_q, p_u)$ on $[0, 1]^2$, which has a unique solution characterized in Lemma 1.

Lemma 1 (Solution to the Unconstrained Employer's Problem)**.** Let p_q, p_u, a^* be defined as above,

and define μ_l, μ_h as follows:

$$\mu_l \equiv \frac{e^{\frac{-x_q}{\lambda}}(e^{\frac{x_u}{\lambda}} - 1)}{e^{\frac{x_u}{\lambda}} - e^{\frac{-x_q}{\lambda}}}, \quad \mu_h \equiv \frac{e^{\frac{x_u}{\lambda}} - 1}{e^{\frac{x_u}{\lambda}} - e^{\frac{-x_q}{\lambda}}}. \tag{7}$$

Then, if $\mu \leq \mu_l$, the employer's unique optimal strategy is to set $p_q^* = p_u^* = 0$, hiring all workers to task zero. If $\mu \geq \mu_h$, the employer's unique optimal strategy is to set $p_q^* = p_u^* = 1$, hiring all workers to task one. Technically, for $\mu = 0$, p_q^* can be arbitrary, but we assume it is 0 for continuity. Similarly, at $\mu = 1$, p_u^* can be arbitrary, but we assume it to be 1. If $\mu \in (\mu_l, \mu_h)$, then the employer's unique optimal strategy is to set p_q^*, p_u^* equal to the following values:

$$p_q^* = \frac{\mu - \mu_l}{\mu(1 - e^{\frac{-x_q}{\lambda}})}, \quad p_u^* = \frac{\mu - \mu_l}{(1-\mu)(e^{\frac{x_u}{\lambda}} - 1)}. \tag{8}$$

Proof. See Appendix 7.1.

So, p_q^* and p_u^* are both increasing in μ on (μ_l, μ_h), but p_q^* is concave and p_u^* is convex. Also, these expressions imply by Bayes' Law that the posteriors corresponding to the high and low signals are equal on the interval (μ_l, μ_h):

$$\begin{aligned} \gamma_h^* \equiv \gamma_\mu(1|p_q^*, p_u^*) &= \frac{\mu p_q^*}{\mu p_q^* + (1-\mu)p_u^*} = \mu_h, \\ \gamma_l^* \equiv \gamma_\mu(0|p_q^*, p_u^*) &= \frac{\mu(1-p_q^*)}{\mu(1-p_q^*) + (1-\mu)(1-p_u^*)} = \mu_l. \end{aligned} \tag{9}$$

3.2 Worker's Problem

Employees with investment cost c invest if and only if their expected net return on investment is weakly greater than c. Given an optimal information structure on the part of the employer, characterized by p_q^* and p_u^*, workers invest if and only if the following condition holds:

$$w \cdot p_q^* - c \geq w \cdot p_u^* \iff w(p_q^* - p_u^*) \geq c.$$

Therefore, the real qualification rate induced among workers of the group subject to the employer's optimal information structure, $\mu^R : [0,1] \to [0,1]$, is

$$\mu^R(\mu) = F(w(p_q^* - p_u^*)). \tag{10}$$

3.3 Unconstrained Equilibrium

As in Coate and Loury (1993), we can then define an equilibrium as a pair of "self-confirming" prior beliefs.

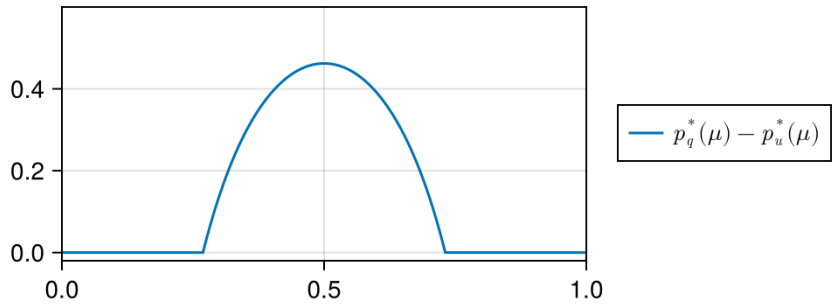

Figure 1: $p_q^*(\mu) - p_u^*(\mu)$, with $x_q = x_u = \lambda = 1$. This figure illustrates the properties that Lemma 2 generalizes to all parameter values: that $p_q^*(\mu) - p_u^*(\mu)$ is continuous, positive-valued, concave, and equal to zero when $\mu \notin (\mu_l, \mu_h)$.

Definition 2 (Unconstrained Equilibrium)**.** An unconstrained equilibrium is a pair of prior beliefs on qualification rate (μ_B^*, μ_W^*) such that, under the best response information structures defined by $(p_{q,B}^*, p_{u,B}^*)$ and $(p_{q,W}^*, p_{u,W}^*)$,

$$\mu^R(\mu_G^*) = \mu_G^*, \quad G = B, W.$$

We call an equilibrium "discriminatory" if $\mu_B^* \neq \mu_W^*$.

Without intervention, the groups are entirely separable and subject to the same parameters x_q, x_u, λ, w. Therefore, discriminatory equilibria are possible if and only if these parameters are such that there are multiple prior beliefs $\mu^* \in (0, 1)$ with $\mu^R(\mu^*) = \mu^*$ ($\mu^* = 0$ is a trivial equilibrium, but we ignore it when considering discriminatory equilibria). Therefore, we will restrict our study of unconstrained equilibria to a single group, as in the employer's problem.

To derive properties of $\mu^R(\mu)$, we must first characterize $p_q^*(\mu) - p_u^*(\mu)$ for $\mu \in [0, 1]$. Figure 1 illustrates $p_q^*(\mu) - p_u^*(\mu)$ under certain parameters. Then, in Lemma 2, we find that $p_q^*(\mu) - p_u^*(\mu)$ has a similar shape with useful properties for all valid x_q, x_u, λ.

Lemma 2. For all $x_q, x_u, \lambda > 0$, we have $0 < \mu_l < \mu_h < 1$, and $p_q^*(\mu) - p_u^*(\mu)$ has the following properties:

(1) $p_q^*(\mu) - p_u^*(\mu)$ is continuous on $[0, 1]$,

(2) $p_q^*(\mu) - p_u^*(\mu) = 0$ for $\mu \leq \mu_l$ and $\mu \geq \mu_h$,

(3) $p_q^*(\mu) - p_u^*(\mu) > 0 \; \forall \; \mu \in (\mu_l, \mu_h)$,

(4) $p_q^*(\mu) - p_u^*(\mu)$ is strictly concave on (μ_l, μ_h).

Proof. See Appendix 7.2.

With this information, I can now characterize the equilibria of this model given particular parameters. There are three possible cases: only the trivial equilibrium of $\mu = 0$, the trivial

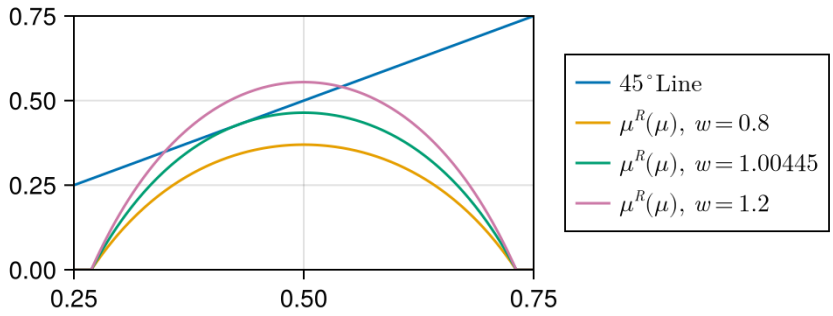

Figure 2: Examples of zero, one, and two equilibria with parameters $x_q = x_u = \lambda = 1, w = 0.8, w \approx 1.00445, w = 1.2$ respectively, and investment cost $c \sim \text{Unif}(0, 1)$. The equilibrium values of μ correspond to the intersections of $\mu^R(\mu)$ with the 45-degree line.

equilibrium and one interior equilibrium, or the trivial equilibrium and two interior equilibria. Figure (2) shows one example of each. Theorem (1) describes necessary and sufficient conditions for the existence of 0, 1, or 2 interior equilibria.

Theorem 1. Consider the value $M = \max_{\mu \in [\mu_l, \mu_h]} (\mu^R(\mu) - \mu)$. Then,

(1) if $M < 0$, there are 0 interior equilibria.

(2) if $M = 0$, there is 1 interior equilibrium.

(3) if $M > 0$, there are 2 interior equilibria.

Additionally, $\mu = 0$ is always a trivial equilibrium.

Proof. By Lemma (2), $p_q^* - p_u^* = 0 \; \forall \; \mu \in [0, \mu_l] \cup [\mu_h, 1]$. Thus, $\mu^R(0) = 0$, so $\mu^* = 0$ is a trivial equilibrium. Additionally, $\mu^R(\mu) - \mu = -\mu < 0$ on $(0, \mu_l] \cup [\mu_h, 1]$, so any further equilibria must occur on (μ_l, μ_h). Since $p_q^* - p_u^*$ and $F(\cdot)$ are continuous, $\mu^R(\mu)$ is continuous on $[\mu_l, \mu_h]$. Therefore, $\mu^R(\mu) - \mu$ attains a maximum on $[\mu_l, \mu_h]$, which we denote M. If this maximum is less than 0, then there is no $\mu \in [\mu_l, \mu_h]$ with $\mu^R(\mu) - \mu = 0$, so there are no non-trivial equilibria.

We are left with the cases where $M \geq 0$. Since $\mu^R(\mu_l) - \mu_l = -\mu_l < 0$ and $\mu^R(\mu_h) - \mu_h = -\mu_h < 0$, any μ such that $\mu^R(\mu) - \mu = M$ is the location of a local maximum. To analyze such μ, we characterize $\frac{d\mu^R}{d\mu}$. For ease of notation, let $d(\mu) = p_q^*(\mu) - p_u^*(\mu)$. Then,

$$\frac{d\mu^R}{d\mu} = F'(w \cdot d(\mu)) \cdot w \cdot d'(\mu). \tag{11}$$

Since F is strictly increasing, $\text{sign}(\frac{d\mu^R}{d\mu}) = \text{sign}(d'(\mu))$. Now, by Lemma (2), $d(\mu)$ is equal to 0 at μ_l and μ_h and strictly concave on the interval between them, so there exists a unique $\mu_c \in (\mu_l, \mu_h)$ with $\frac{d}{d\mu}(d(\mu)) = 0$. $d(\mu)$ is strictly increasing on (μ_l, μ_c), and $d(\mu)$ is strictly decreasing on (μ_c, μ_h).

Now, we will characterize the behavior of the second derivative of μ^R on (μ_l, μ_c). Since $F'(\cdot) > 0$ and $d'(\mu) > 0$ on this interval, by differentiating and rearranging the expression we find

that

$$\text{sign}\left(\frac{d^2\mu^R}{d\mu^2}\right) = \text{sign}\left(\frac{F''(w\cdot d(\mu))}{F'(w\cdot d(\mu))} + \frac{d''(\mu)}{w\cdot d'(\mu)^2}\right). \tag{12}$$

Since F is the CDF of a log-concave distribution, F' is log-concave. This means that $\frac{F''(w\cdot d(\mu))}{F'(w\cdot d(\mu))}$ is non-increasing in its argument. Since $d(\mu)$ is increasing on the interval we are currently considering, $\frac{F''(w\cdot d(\mu))}{F'(w\cdot d(\mu))}$ is non-increasing in μ. Then, we have

$$\frac{d}{d\mu}\left(\frac{d''(\mu)}{w\cdot d'(\mu)^2}\right) = \frac{d'(\mu)d'''(\mu) - 2d''(\mu)^2}{w\cdot d'(\mu)^3}.$$

A direct computation shows this equals a negative expression for all $\mu \in (\mu_l, \mu_c)$; details available upon request.

Thus, $\frac{F''(w\cdot d(\mu))}{F'(w\cdot d(\mu))} + \frac{d''(\mu)}{w\cdot d'(\mu)^2}$ is strictly decreasing on (μ_l, μ_c). Since $d'(\mu_c) = 0, F'(\cdot) > 0$, and $d''(\mu) < 0$ on (μ_l, μ_h), we know that $\frac{d^2\mu^R}{d\mu^2} < 0$ at μ_c, and thus it is also negative in a small radius of μ_c by continuity. Therefore, $\frac{F''(w\cdot d(\mu))}{F'(w\cdot d(\mu))} + \frac{d''(\mu)}{w\cdot d'(\mu)^2} < 0$ in that same small radius of μ_c. This implies that $\frac{F''(w\cdot d(\mu))}{F'(w\cdot d(\mu))} + \frac{d''(\mu)}{w\cdot d'(\mu)^2}$ is either strictly negative on (μ_l, μ_c), or that it starts positive, crosses 0 precisely once, then is negative after that. Consequently, by Equation (12), the second derivative of μ^R has the same property. We thus know that if the second derivative of μ^R is weakly negative at some $\hat{\mu} \in (\mu_l, \mu_c)$, it is strictly negative for all $\mu \in (\hat{\mu}, \mu_c)$.

Then, since M is attained at an interior point, we have

$$\arg\max_{\mu\in[\mu_l,\mu_h]} (\mu^R(\mu) - \mu) \subseteq \mathcal{M} := \left\{\mu : \frac{d\mu^R}{d\mu}(\mu) = 1,\ \frac{d^2\mu^R}{d\mu^2}(\mu) \leq 0\right\}, \tag{13}$$

and $\mathcal{M}$ is not empty.

We now seek to show that if $M \geq 0$, then $\mathcal{M}$ is a singleton. Let $\mu^* \in \mathcal{M}$ be arbitrary. First, since $\frac{d\mu^R}{d\mu}(\mu) = 1 > 0$ on $\mathcal{M}$, we know that $\mathcal{M} \subseteq [\mu_l, \mu_c)$. Then, as we just established, $\frac{d^2\mu^R}{d\mu^2}(\mu^*) \leq 0$ implies that $\frac{d^2\mu^R}{d\mu^2}(\mu) < 0$ for all $\mu \in (\mu^*, \mu_c)$. So, $\frac{d\mu^R}{d\mu}(\mu)$ is strictly decreasing on (μ^*, μ_c), which implies that $\frac{d\mu^R}{d\mu}(\mu) < 1$ on (μ^*, μ_c) and in turn that $\mathcal{M} \cap (\mu^*, \mu_c) = \emptyset$. Therefore, μ^* is the maximum of $\mathcal{M}$. Since μ^* was arbitrary, though, and $\mathcal{M}$ is non-empty, we conclude that μ^* is the only element in $\mathcal{M}$.

When $M = 0$, all interior equilibria are in $\mathcal{M}$, since equilibria are points where $\mu^R(\mu) - \mu = M = 0$ by definition, so we conclude that there is precisely one equilibrium. What remains is $M > 0$. Let μ^* be the single element of $\mathcal{M}$. Note that since $\mu^R(\mu_l) - \mu_l < 0$ and $\mu^R(\mu_h) - \mu_h < 0$, the Intermediate Value Theorem implies there is at least one point on (μ_l, μ^*) and one point on (μ^*, μ_h) where $\mu^R(\mu) - \mu = 0$. Thus, there are at least two interior equilibria. As established before, $\frac{d\mu^R}{d\mu}(\mu) < 1$ on (μ^*, μ_c) and $\frac{d\mu^R}{d\mu}(\mu) \leq 0$ on (μ_c, μ_h), so $\mu^R(\mu) - \mu$ is strictly decreasing on (μ^*, μ_h). Thus, precisely one equilibrium occurs on (μ^*, μ_h).

Let $\underline{\mu}$ be the smallest interior equilibrium: $\underline{\mu} = \min\{\mu : \mu^R(\mu) - \mu = 0\}$. Note that

$\underline{\mu} \in (\mu_l, \mu^*)$ since we established at least one equilibrium exists in this interval. Since

$$\mu^R(\mu_l) - \mu_l < 0 = \mu^R(\underline{\mu}) - \underline{\mu} < \mu^R(\mu^*) - \mu^* = M,$$

the Mean Value Theorem implies that there exists a point $\hat{\mu} \in (\mu_l, \underline{\mu})$ with $\frac{d\mu^R}{d\mu}(\hat{\mu}) > 1$. Now, assume for contradiction there exists a point $\bar{\mu} \in (\hat{\mu}, \mu^*)$ with $\frac{d\mu^R}{d\mu}(\bar{\mu}) \leq 1$. The Mean Value Theorem implies that there is a point on $(\hat{\mu}, \bar{\mu})$ with a negative second derivative of μ^R. Using our earlier result that the second derivative, once negative, remains strictly negative, this in turn implies that the second derivative of μ^R is strictly negative for all μ greater than that point. In particular, it is negative on $(\bar{\mu}, \mu^*)$, and thus we have $\frac{d\mu^R}{d\mu}(\mu*) < \frac{d\mu^R}{d\mu}(\bar{\mu}) \leq 1$, which is a contradiction. Therefore, $\frac{d\mu^R}{d\mu}(\mu) > 1$ on $(\hat{\mu}, \mu^*)$. This implies that $\mu^R(\mu) - \mu$ is strictly increasing on $(\underline{\mu}, \mu^*) \subset (\hat{\mu}, \mu^*)$, and thus that $\mu^R(\mu) - \mu > 0$ on $(\underline{\mu}, \mu^*)$. So, $\underline{\mu}$ is the only equilibrium on (μ_l, μ^*), and we conclude there are precisely two equilibria. □

Thus, discriminatory equilibria exist whenever $M > 0$. Importantly, as in Coate and Loury (1993), the employers in the model correctly recognize that qualification rates are disparate between groups and act accordingly, but this disparity cannot be attributed to some exogenous difference between the groups. The gaps only exist because the employers expect them to exist.

4 Affirmative Action

In this section, I implement three versions of "affirmative action" policies in the form of constraints on the information structure selected by employers. For each, I derive an updated solution to the employer's problem then characterize whether each is effective at eliminating discrimination.

4.1 Information Requirement for Assignment to Task Zero

The first intervention I employ is a policy preventing employers from assigning a worker to task zero unless they are sufficiently sure that the worker is not qualified. This policy can be modeled by a constraint on the low posterior, $\gamma_u \equiv \gamma_\mu(0|p_q, p_u) \leq \gamma_c$, where a stronger version of the policy corresponds to a lower γ_c. In the real world, such a policy could look like a requirement that a greater share of applicants receive interviews, such that only those most obviously unqualified are rejected based solely on their résumé.

As established in Equation (9), in the unconstrained case, for $\mu \in [\mu_l, \mu_h]$, $\gamma_l = \mu_l$, and $\gamma_h = \mu_h$. Therefore, for this constraint to be binding on the interval on which non-trivial equilibria can occur (and thus have the potential to prevent discrimination), γ_c must be strictly less than μ_l. When this is the case, the employer's optimal strategy will be altered for all priors $\mu \in (\gamma_c, \mu_h)$. Below γ_c, the original strategy of assigning all workers to task zero is still allowable since the posterior (which is equal to the prior) is less than γ_c. Above μ_h, no one is ever rejected, so the original strategy of assigning all workers to task one is still allowable.

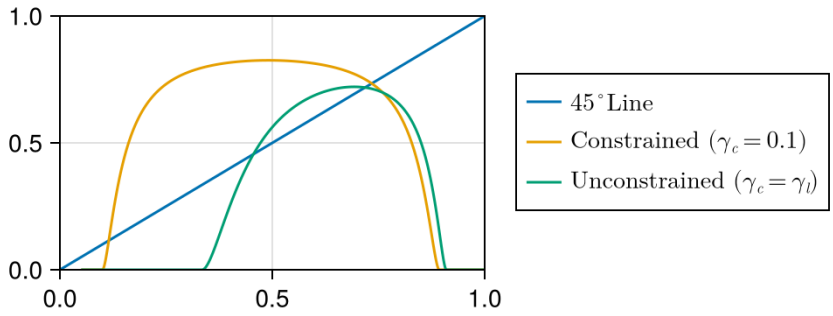

Figure 3: Realized qualification rate $\mu^R(\mu, \gamma_c)$, a function of prior μ and policy γ_c. In the constrained case, $\gamma_c = 0.1$, and the unconstrained case is $\gamma_c = \mu_l$. The parameters are set to $x_q = 1, x_u = 2, \lambda = 1, w = 4$ and investment cost $c \sim \text{Gamma}(2, 1)$. Equilibria are intersections of $\mu^R(\mu, \gamma_c)$ with the 45-degree line.

By the concavity of the objective function, when the constraint binds, γ_l will equal γ_c under the employer's optimal strategy. This defines a precise relationship between p_q^* and p_u^* in optimum for $\mu \in (\gamma_c, \mu_h)$:

$$p_q^* = \frac{\gamma_c(1-\mu)\, p_u^* + \mu - \gamma_c}{(1-\gamma_c)\mu}. \tag{14}$$

Then, $R(p_q^*(p_u), p_u)$ is smooth and strictly concave with respect to p_u (for any $\mu \in (0, 1)$, so by taking the derivative, setting the expression equal to zero, and substituting in the above equality, we have the following implicit equation defining the unique optimal p_u^*:

$$\log\left(1 + \frac{\mu - \gamma_c}{p_u(1-\mu)}\right) + \gamma_c \log\left(\frac{\gamma_c(1-\mu)p_u}{\gamma_c(1-\mu)p_u + \mu - \gamma_c}\right) = \frac{x_u(1-\gamma_c) - x_q\gamma_c}{\lambda}. \tag{15}$$

For all valid parameters, this defines a $p_u^* \geq 0$; if the p_u^* that satisfies this equation is greater than 1, then $p_u^* = p_q^* = 1$, as this implies that the derivative of R with respect to p_u is positive for all valid $p_u \in [0, 1]$. Based on this expression, we can numerically approximate the employer's optimal strategy given values for $\mu, \gamma_c, x_q, x_u, \lambda$. Figure (3) shows an example of the effect of this affirmative action policy on the realized qualification rate $\mu^R(\mu, \gamma_c)$.

In this example, affirmative action fails to alleviate discrimination. In fact, in this case, the intervention made the two equilibria further apart from one another. It turns out that this failure occurs in all cases where discrimination is possible without affirmative action. To demonstrate this, I first prove the following lemmas, beginning with characterizing a set of useful properties of $p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c)$:

Lemma 3. $p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c)$ has properties analogous to Lemma (2): for a fixed γ_c, the function is continuous on $[0, 1]$, strictly positive and concave on some open interior interval, and equal to zero outside of that interval. Furthermore, $p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c)$ is jointly continuous on the domain $(\mu, \gamma_c) \in [0, 1] \times (0, \mu_l]$, and it is differentiable at all points (μ, γ_c) satisfying $p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c) > 0$.

Proof. See Appendix 7.3.

The next lemma mathematically formalizes the sense in which by decreasing γ_c, implementing a stricter affirmative action policy, you "stretch" the curve $p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c)$ up and to the left. It asserts that the maximum of this curve is weakly increasing in γ_c and that all points to the left of that maximum at a given γ_c are decreasing in γ_c (increasing as γ_c gets lower).

Lemma 4.

1. For all μ such that $\dfrac{\partial}{\partial \mu}(p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c)) > 0$, that is, all μ on the increasing half of the curve at a fixed γ_c, $\dfrac{\partial}{\partial \gamma_c}(p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c)) < 0$.
2. For $\gamma_c \in (0, \mu_l]$, $m(\gamma_c) = \max\limits_{\mu \in [\gamma_c, \mu_h]}(p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c))$ is decreasing in γ_c.

Proof. Proof combines analytical arguments with numerical verification of certain derivative conditions across the parameter space; see Appendix 7.3.

Based on these two lemmas, we can show that in any case where there are multiple equilibria without discrimination, there exists no affirmative action policy γ_c under which discrimination is alleviated. This result is stated and proven in Proposition (1).

Proposition 1. When discrimination is feasible without intervention, an affirmative action policy that demands sufficient confidence that a worker is unqualified before assignment to task zero, requiring $\gamma_l \leq \gamma_c \in (0, \mu_l)$, cannot prevent discrimination for any such γ_c.

Proof. See Appendix 7.3.

As an example, consider a secretary position in which the primary skill required is typing ability and where the two groups in consideration are those from affluent or poorer backgrounds. An employer for such a position might believe that those from an affluent background are more likely to be competent at typing, even if access to computers was not so disparate in reality, giving them the benefit of the doubt as compared to those from poorer backgrounds. Perhaps the typing test for those from poorer backgrounds is *so* difficult that even the qualified usually fail, and the test for those from more affluent backgrounds is much easier. When this affirmative action is applied, the test for workers of poor background must change to reduce the rate of qualification among those rejected—perhaps they make the test slightly easier, so that a much greater share of the qualified succeed. Now, at any relatively low prior qualification level, the returns to qualification increase. As a result, a prior even lower than the original low equilibrium, at which the testing used to be so uninformative that the realized qualification rate was lower still, now becomes the new low equilibrium as the incentive for workers to invest given the employer's response to this prior has increased. This version of affirmative action simply exacerbated the issue.

This result demands a deep caution on the part of policy-makers. This sort of affirmative action policy, while nobly aimed to prevent unfairly imprecise signals being used for groups with

low qualification levels, may harm the very group it intends to help by lowering the qualification rate of the low equilibrium. When designing programs intending to prevent rejection of qualified candidates from groups with lower aggregate qualification rates, policy-makers must be careful that the new equilibrium reached under the policy does not resolve to one with even worse outcomes than before the policy was implemented. My static model does not provide insight as to how the shift from unconstrained to constrained equilibrium occurs, but the unavoidable presence of multiple equilibria after affirmative action is applied is concerning nonetheless.

4.2 Information Requirement for Assignment to Task One

The second intervention I examine is preventing employers from assigning a worker to task one unless the employer is sufficiently sure that the worker is qualified. Formally, this is a constraint on the high posterior, $\gamma_h \equiv \gamma_\mu(1|p_q, p_u) \geq \gamma_c$, where a stronger version of the policy corresponds to a higher γ_c. An example of this policy may be requiring a greater number of rounds of interviews before any candidate is hired.

Similarly to the information requirement for assignment to task zero, Equation (9) implies that for this constraint to be binding on the interval on which non-trivial equilibria can occur, γ_c must be strictly greater than μ_h. When this is satisfied, the employer's optimal information structure changes for all priors $\mu \in (\mu_l, \gamma_c)$. Below μ_l, nobody is assigned to task one so the constraint never applies. Above γ_c, everyone is assigned to task one and the posterior is equal to the prior, greater than γ_c, so the constraint is satisfied.

Similarly to the constraint on the low posterior, when the constraint binds, it will hold with equality by the concavity of the objective function. Then $\gamma_h = \gamma_c$ implies the following relationship between p_q^* and p_u^*:

$$p_q^* = \frac{\gamma_c(1-\mu)}{(1-\gamma_c)\mu} p_u^* \tag{16}$$

Then, once again, differentiating $R(p_q^*(p_u), p_u)$ with respect to p_u and setting it equal to zero, we have an implicit expression defining p_u^*.

$$\log\left(1 - \frac{1-\mu}{1-\gamma_c}p_u^*\right) - \gamma_c \log\left(\frac{\mu}{\gamma_c} - \frac{1-\mu}{1-\gamma_c}p_u^*\right) - (1-\gamma_c)\log\left(\frac{1-\mu}{1-\gamma_c}(1-p_u^*)\right) = \frac{x_q\gamma_c - x_u(1-\gamma_c)}{\lambda} \tag{17}$$

For all valid parameters, this defines a $p_u^* \leq 1$; if the p_u^* that satisfies this equation is less than 0, then $p_u^* = p_q^* = 0$. As before, we can numerically approximate the employer's optimal strategy for particular parameters. Figure (4) shows a promising example of this version of affirmative action successfully eliminating discriminatory equilibria for certain parameters.

Unfortunately, this success is not possible in all cases where discriminatory equilibria exist without intervention. Figure (5) graphs $\max_{\mu \in [\mu_l, \gamma_c]} (\mu^R(\mu, \gamma_c) - \mu)$ as a function of γ_c. For values of γ_c where this function is positive, there are multiple equilibria after affirmative action is applied, since μ^R_{const} is continuous and $\mu^R_{\text{const}}(\mu_l) = \mu^R_{\text{const}}(\gamma_c) = 0$.

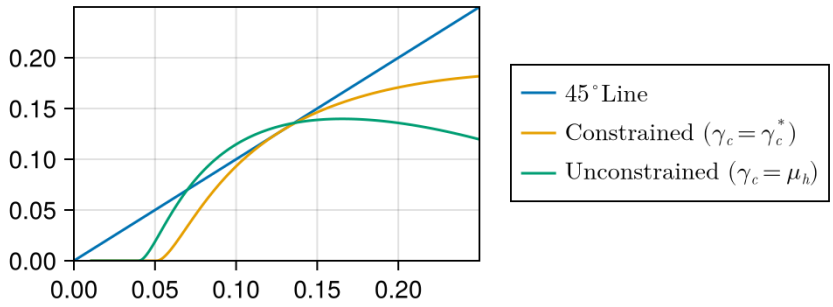

Figure 4: Realized qualification rate μ^R as a function of prior μ with and without an affirmative action policy of $\gamma_c = 0.7843$, with parameters $x_q = 4, x_u = 1, \lambda = 1.6, w = 1$ and investment cost $c \sim \text{Gamma}(2, 1)$. Equilibria are intersections with the 45-degree line.

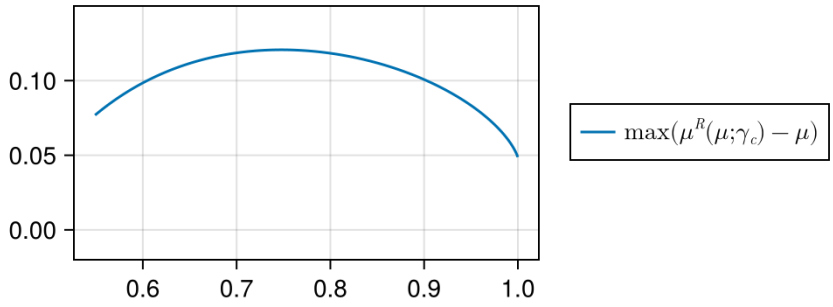

Figure 5: Maximum distance from the 45-degree line, $\mu^R(\mu, \gamma_c) - \mu$ as a function of intervention strength γ_c, with parameters $x_q = 2, x_u = 1, \lambda = 1.6, w = 3$ and investment cost $c \sim \text{Gamma}(2, 1)$. Values greater than 0 imply multiple interior equilibria under the affirmative action intervention.

As seen in the figure, since this function takes strictly positive values on the entire domain $(\mu_h, 1)$, there exists no intervention $\gamma_c \in (\mu_h, 1)$ that prevents affirmative action. In fact, there are cases in which this affirmative action restriction not only fails to prevent discrimination, but actually makes the gap between low and high equilibria larger than the unconstrained case, for all values of $\gamma_c \in (\mu_h, 1)$. Figure (6) shows an example of this, plotting the gap between low and high interior equilibria as a function of γ_c.

In this example, affirmative action makes the qualification rate gap between equilibria strictly worse than before. In this example, on most of the domain of possible γ_c values, increasing the strength of the affirmative action restriction by increasing γ_c makes the gap wider and wider. So, while this version of affirmative action can alleviate discrimination, depending on the specifics of equilibria pre-intervention, it may have the opposite of its intended effect. Proposition (2) formalizes a sufficient condition for the success of this type of affirmative action.

Proposition 2. Let discrimination be possible without affirmative action, with $M > 0$ according to Theorem (1). Allow the technical assumption that $e^{\frac{x_q}{\lambda}} > w\frac{F''(w)}{2F'(w)} + 1$. Note that when w lies on the decreasing side of F' (recall that F' is log-concave), this is always satisfied. Define

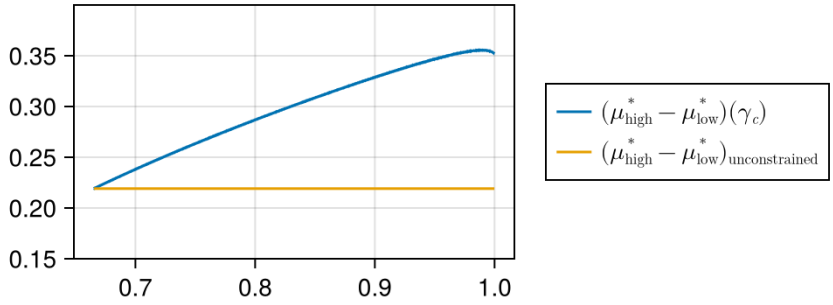

Figure 6: The gap between equilibria, $\mu^*_{\text{high}} - \mu^*_{\text{low}}$, as a function of γ_c. The orange line is the gap between equilibria without affirmative action for comparison.

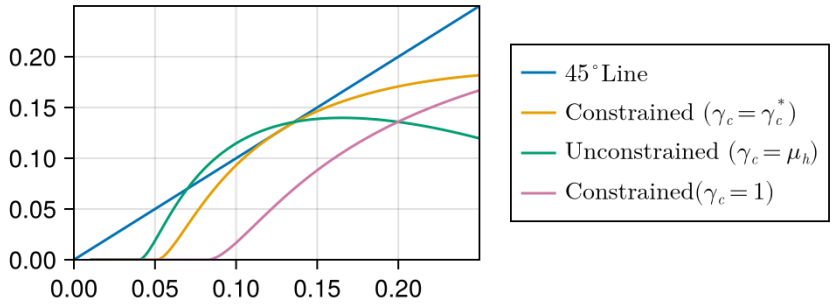

Figure 7: Realized qualification rate μ^R as a function of prior μ without affirmative action and with affirmative action policies of $\gamma_c = 0.7843, 1$ and parameters $x_q = 4, x_u = 1, \lambda = 1.6, w = 1$ with investment cost $c \sim \text{Gamma}(2, 1)$. Equilibria are intersections with the 45-degree line.

$M(\gamma_c) = \max_{\mu \in [\mu_l, \gamma_c]} (\mu^R(\mu, \gamma_c) - \mu)$, so that $M(\mu_h)$ is equal to M as defined in Theorem (1).

Let $\gamma_c = 1$, which is outside the normal allowable bounds, but makes the employer's best response computable in closed form (p^*_u must equal 0 for $\mu < 1$, and is assumed to be 0 at $\mu = 1$ for continuity). If $M(1) < 0$, then there exists some $\gamma_c \in (\mu_h, 1)$ under which there is precisely one interior equilibrium, where discrimination is not possible.

Proof. Proof combines analytical arguments with numerical verification of one derivative condition; see Appendix 7.3. □

This sufficient condition for the success of some affirmative action policy captures the intuition that "if the strongest possible affirmative action policy is overkill, then there is some intermediate policy that will alleviate discrimination." An illustration can be seen in Figure (7), which uses the same parameters as the successful example seen in Figure (4) but now includes the curve $\mu^R(\mu, 1)$. This curve lies entirely below the 45-degree line, and thus, as per the proposition, an intermediate value of γ_c alleviates discrimination.

As an example, consider the same secretary position from before, in which employers discriminate against those from poorer backgrounds based on perceptions about typing ability. Imag-

ine an affirmative action policy requiring that essentially *zero* unqualified people are ever hired, and that the cost of making a very precise test is great enough that employers instead make tests so hard that almost everyone fails, qualified and unqualified. Now, based on the employer's best response strategy given *any* prior belief on qualification rate, the incentive for workers to invest is low enough that the realized qualification rate never meets the prior. This proposition asserts that some intermediate value that is less restrictive on the employers will result in tests being just precise enough that equilibrium occurs at exactly one qualification rate.

4.3 Equal Opportunity for the Qualified

The final intervention I will analyze is a requirement that qualified individuals have the same probability of being assigned to task one regardless of their group. Formally, this is the constraint that $p_{q,B} = p_{q,W} = p_q$. This sort of policy may be considered the moral "ideal," but it would certainly be the hardest to enforce in practice. Nevertheless, it is still worth examining, as although the probability of returning a high signal is the same for the qualified in both groups, that probability for the unqualified may be chosen freely, so it is not immediately obvious that this policy will prevent all discriminatory equilibria.

With a share s of workers in the population being of group B, the employer's objective function is as follows:

$$R_{\text{full}}(p_q, p_{u,B}, p_{u,W}) = sR(p_q, p_{u,B}|\mu_B) + (1-s)R(p_q, p_{u,W}|\mu_W) \tag{18}$$

Under this policy, the employer can be seen as carrying out their optimization in two stages. Given an arbitrary p_q, along with their prior beliefs μ_B, μ_W, they solve for the optimal $p_{u,B}$ and $p_{u,W}$, since the groups are separable at this point. Then, given the function mapping p_q and μ_G to $p_{u,G}$, they can maximize their objective function solely as a function of p_q. To solve for $p^*_{u,G}$ as a function of p_q and μ_G, we simply set the partial derivative of R with respect to p_u equal to zero, yielding the following expression:

$$p^*_{u,G}(p_q, \mu_G) = \frac{\mu + (e^{\frac{x_u}{\lambda}} - 1)(1 - \mu p_q) - \sqrt{(\mu + (e^{\frac{x_u}{\lambda}} - 1)(1 - \mu p_q))^2 - 4(e^{\frac{x_u}{\lambda}} - 1)(1-\mu)\mu p_q}}{2(e^{\frac{x_u}{\lambda}} - 1)(1-\mu)} \tag{19}$$

It holds that $p^*_{u,G}$ is increasing in μ for a fixed p_q. So, for any p_q, if $\mu_W > \mu_B$, $p^*_{u,W}(p_q, \mu_W) > p^*_{u,B}(p_q, \mu_B)$. This implies that $w(p_q - p^*_{u,W}(p_q, \mu_W)) < w(p_q - p^*_{u,B}(p_q, \mu_B))$, and since F is increasing, that $F(w(p_q - p^*_{u,W}(p_q, \mu_W))) \leq F(w(p_q - p^*_{u,B}(p_q, \mu_B)))$. Thus, without ever solving for the optimal p_q itself, this result proves that there can be no equilibria where $\mu_B \neq \mu_W$, since anytime $\mu_B < \mu_W$, the realized qualification rates will satisfy $\mu^R_B \geq \mu^R_W$, and anytime $\mu_B > \mu_W$, we have $\mu^R_B \leq \mu^R_W$.

When $\mu_B = \mu_W$, the employer's problem is identical to the unconstrained case, so for any unconstrained equilibrium qualification rate μ^*, $\mu_B = \mu_W = \mu^*$ is an equilibrium under this version of affirmative action. These results are stated in Proposition (3).

Proposition 3. The policy of equal opportunity for the qualified, that is, $p_{q,B} = p_{q,W}$, preserves all equilibria from the unconstrained case but precludes separation by group between them, preventing discriminatory equilibria.

The intuition behind this result is that in general, the employer can stomach accidentally hiring a greater share of unqualified individuals within a group when that group's aggregate qualification rate is higher. When the employer is required to give equal opportunity to the qualified in both groups, this changes their strategy such that the information structure by which they assess those in the less qualified group is more "accurate," since they have a necessity to hire a lesser share unqualified individuals from the group that is less qualified overall. Thus, the incentives are inverted: those from the less qualified group stand to gain more by getting qualified since they are subject to a more accurate test, so their realized qualification rate is higher. Thus, no such disparity in qualification rates is possible in equilibrium.

As stated above, this version of affirmative action is certainly most difficult to implement, but this strong positive result provides some hope that policies which aim to implement this constraint can have positive effects toward reducing or eliminating statistical discrimination.

5 Conclusion

Especially in the aftermath of the US Supreme Court's ban on race-based affirmative action in higher education ("Students for Fair Admissions, Inc. v. President and Fellows of Harvard College", 2023), increased scrutiny has been placed on all affirmative action regimes and alternatives to the strict "extra points" versions of affirmative action formerly seen in many college admissions processes will continue to be explored. This model provides sharp results on affirmative action policies regulating information acquisition. A policy requiring confidence before rejection always fails, a policy require confidence before hiring has mixed effects, and mandating equal opportunity for the qualified always succeeds. The difference lies in how they shape incentives: the first policy increases investment incentives at low priors, inducing equilibria at even lower qualification rates, whereas the last inverts the employer's incentive asymmetry entirely. Of course, the model has some limitations: it does not provide for convex costs of hiring to Task One, a paradigm which typically exists in the real world, often in its limiting case of fixed capacity; additionally, it follows Coate and Loury's (1993) assumption of exogenous wages.

This paper provides a renewal of the warning that Coate and Loury (1993) offer to policy-makers, this time regarding a newer, subtler version of affirmative action. Policies intended to promote the acquisition of more accurate information by employers while hiring may under certain circumstances hurt the very groups that they were intended to help. Policy-makers must therefore pay close attention to the specifics of any labor-market in which they may apply such a policy, carefully considering how their intervention alters incentives to avoid falling into this trap.

6 References

Arcidiacono, P., Aucejo, E. M., Fang, H., & Spenner, K. I. (2011). Does affirmative action lead to mismatch? A new test and evidence. *Quantitative Economics*, *2*(3), 303–333. https://doi.org/10.3982/QE83

Arrow, K. J. (1995). The theory of discrimination [Originally issued as Working Paper 30A, Princeton University Industrial Relations Section]. In O. Ashenfelter & K. Hallock (Eds.), *Labor economics* (Vol. 4). Elgar. http://arks.princeton.edu/ark:/88435/dsp014t64gn18f (Original work published 1971)

Bleemer, Z. (2022). Affirmative Action, Mismatch, and Economic Mobility after California's Proposition 209*. *The Quarterly Journal of Economics*, *137*(1), 115–160. https://doi.org/10.1093/qje/qjab027

Coate, S., & Loury, G. C. (1993). Will Affirmative-Action Policies Eliminate Negative Stereotypes? *The American Economic Review*, *83*(5), 1220–1240. Retrieved October 3, 2024, from http://www.jstor.org/stable/2117558

Echenique, F., & Li, A. (2025). Rationally inattentive statistical discrimination: Arrow meets phelps. *Journal of the European Economic Association*, *23*(5), 1712–1742. https://doi.org/10.1093/jeea/jvaf002

Fang, H., & Moro, A. (2011, January). Chapter 5 - Theories of Statistical Discrimination and Affirmative Action: A Survey. In J. Benhabib, A. Bisin, & M. O. Jackson (Eds.), *Handbook of Social Economics* (pp. 133–200, Vol. 1). North-Holland. https://www.sciencedirect.com/science/article/pii/B978044453187200005X

Matějka, F., & McKay, A. (2015). Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model. *The American Economic Review*, *105*(1), 272–298. Retrieved November 11, 2024, from https://www.jstor.org/stable/43497060

Students for fair admissions, inc. v. president and fellows of harvard college. (2023, June 29). Retrieved April 14, 2025, from https://supreme.justia.com/cases/federal/us/600/20-1199/

7 Appendix

7.1 Proofs for Section 3.1

PROOF OF LEMMA 1: As argued by Matějka and McKay (2015), the concavity of the Shannon entropy function implies the convexity of the cost function, which in turn implies the concavity of the employer's objective function. Then, since R is continuous on $[0,1]^2$ and bounded above by its concavity, it attains a maximum on that domain. There are two possible cases for this maximum.

First, by the concavity and continuity of R, if there exists only one local maximum of R on $(0,1)^2$, then it is the unique maximum of R on the closed domain. If there does not exist a local max of R on $(0,1)^2$, the maximum must be on the boundary, and we will leverage the Karush-Kuhn-Tucker conditions to identify it and demonstrate uniqueness.

The first order conditions of $R(p_q, p_u)$ on $(0,1)^2$ are as follows:

$$\frac{\partial}{\partial p_q} R(p_q,p_u) = x_q\mu + \lambda\mu \log\left(\frac{(1-p_q)(\mu p_q + (1-\mu)p_u)}{p_q(1-(\mu p_q+(1-\mu)p_u))}\right) = 0$$

$$\frac{\partial}{\partial p_u} R(p_q,p_u) = -x_u(1-\mu) + \lambda(1-\mu) \log\left(\frac{(1-p_u)(\mu p_q + (1-\mu)p_u)}{p_u(1-(\mu p_q+(1-\mu)p_u))}\right) = 0$$

Which induce the following expressions for p_q^* and p_u^*:

$$p_q^* = \frac{1}{1-e^{\frac{-x_q}{\lambda}}}\left(1 - \frac{e^{\frac{-x_q}{\lambda}}(e^{\frac{x_u}{\lambda}}-1)}{\mu(e^{\frac{x_u}{\lambda}} - e^{\frac{-x_q}{\lambda}})}\right), \quad p_u^* = \frac{1}{1-\mu}\left(\frac{\mu}{e^{\frac{x_u}{\lambda}}-1} - \frac{e^{\frac{-x_q}{\lambda}}}{e^{\frac{x_u}{\lambda}} - e^{\frac{-x_q}{\lambda}}}\right).$$

The first order conditions also define the following relationship between p_q^* and p_u^*:

$$p_u^* = \frac{e^{\frac{-x_q}{\lambda}} p_q^*}{e^{\frac{-x_q}{\lambda}} p_q^* + e^{\frac{x_u}{\lambda}}(1-p_q^*)},$$

which is a bijection from $(0,1) \to (0,1)$. Therefore any necessary and sufficient conditions for $p_q^* \in (0,1)$ also ensure $p_u^* \in (0,1)$. The following two conditions ensure that $p_q^* > 0$ and $p_q^* < 1$ respectively:

$$\frac{e^{\frac{-x_q}{\lambda}}(e^{\frac{x_u}{\lambda}}-1)}{\mu(e^{\frac{x_u}{\lambda}} - e^{\frac{-x_q}{\lambda}})} < 1, \quad \frac{e^{\frac{x_u}{\lambda}}-1}{\mu(e^{\frac{x_u}{\lambda}} - e^{\frac{-x_q}{\lambda}})} > 1,$$

or equivalently,

$$\mu_l \equiv \frac{e^{\frac{-x_q}{\lambda}}(e^{\frac{x_u}{\lambda}}-1)}{e^{\frac{x_u}{\lambda}} - e^{\frac{-x_q}{\lambda}}} < \mu < \mu_h \equiv \frac{e^{\frac{x_u}{\lambda}}-1}{e^{\frac{x_u}{\lambda}} - e^{\frac{-x_q}{\lambda}}}.$$

Therefore, for $\mu \in (\mu_h, \mu_l)$, R is uniquely maximized at (p_q, p_u) as defined above, and for $\mu \notin (\mu_h, \mu_l)$, there is no local maximum of R on $(0,1)^2$.

For $\mu \notin (\mu_l, \mu_h)$, we use the Karush-Kuhn-Tucker conditions which are necessary for optimality. Our constraints are $-p_q \le 0, -p_u \le 0, p_q - 1 \le 0, p_u - 1 \le 0$, and we wish to minimize

$-R(p_q, p_u)$. Therefore our conditions are:

1. $-\frac{dR}{dp_q} - \lambda_1 + \lambda_2 = 0,\ -\frac{dR}{dp_u} - \lambda_3 + \lambda_4 = 0,$
2. $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \geq 0,$
3. $\lambda_1 p_q = 0,\ \lambda_2(p_q - 1) = 0,\ \lambda_3 p_u = 0,\ \lambda_4(p_u - 1) = 0$
4. $-p_q \leq 0,\ -p_u \leq 0,\ p_q - 1 \leq 0,\ p_u - 1 \leq 0.$

We know that in this case there exist no local minima on the interior domain $(0,1)^2$, so we can ignore that domain from consideration since the conditions cannot be satisfied. At points where $p_q = 0$ or $p_q = 1$, besides $(0,0)$ and $(1,1)$, $-\frac{dR}{dp_q} = \infty$. At points where $p_u = 0$ or $p_u = 1$, besides $(0,0)$ and $(1,1)$, $-\frac{dR}{dp_u} = \infty$. Therefore none of these points are candidates for the location of a global minimum, as the conditions cannot be satisfied there, either. What remain are $(0,0)$ and $(1,1)$. Since a maximum of R must exist on $[0,1]^2$, we can simply select the point that induces a greater value of R.

$R(0,0) > R(1,1)$ if and only if $\mu < \frac{x_u}{x_q + x_u}$, and $R(0,0) < R(1,1)$ if and only if $\mu > \frac{x_u}{x_q + x_u}$. We now show $\mu_l < \frac{x_u}{x_q + x_u}$, and $\mu_h > \frac{x_u}{x_q + x_u}$, which then implies that for $\mu \leq \mu_l$, $(0,0)$ is the location of the unique maximum of R, and for $\mu \geq \mu_h$, $(1,1)$ is the location of the unique maximum of R. Beginning with μ_l,

$$\mu_l = \frac{e^{\frac{-x_q}{\lambda}}(e^{\frac{x_u}{\lambda}} - 1)}{e^{\frac{x_u}{\lambda}} - e^{\frac{-x_q}{\lambda}}} < \frac{x_u}{x_q + x_u} \iff \frac{e^{\frac{-x_q}{\lambda}}}{1 - e^{\frac{-x_q}{\lambda}}} x_q < \frac{e^{\frac{x_u}{\lambda}}}{e^{\frac{x_u}{\lambda}} - 1} x_u.$$

The left hand side of this inequality is decreasing in x_q, and the limit as $x_q \to 0$ is λ. The right hand side of this inequality is increasing in x_u, and the limit as $x_u \to 0$ is also λ. Therefore, for $x_q, x_u > 0$, $\mu_l < \frac{x_u}{x_q + x_u}$. Next,

$$\mu_h = \frac{e^{\frac{x_u}{\lambda}} - 1}{e^{\frac{x_u}{\lambda}} - e^{\frac{-x_q}{\lambda}}} > \frac{x_u}{x_q + x_u} \iff \frac{x_q}{1 - e^{\frac{-x_q}{\lambda}}} > \frac{x_u}{e^{\frac{x_u}{\lambda}} - 1}.$$

The left hand side of this inequality is increasing in x_q, the limit as $x_q \to 0$ is λ, and the right hand side is decreasing in x_u, with the limit as $x_u \to 0$ also being λ. Thus, for $x_q, x_u > 0$, $\mu_h > \frac{x_u}{x_q + x_u}$.

Therefore, for $\mu \in (\mu_l, \mu_h)$, the employer's unique optimal strategy is selection of p_q^*, p_u^* as defined above, for $\mu \leq \mu_l$, their optimal strategy is $p_q^* = p_u^* = 0$, and for $\mu \geq \mu_h$, their optimal strategy is $p_q^* = p_u^* = 1$. □

7.2 Proofs for Section 3.3

PROOF OF LEMMA 2:

Proving $0 < \mu_l < \mu_h < 1$: first, for $x_q, x_u, \lambda > 0, e^{\frac{-x_q}{\lambda}} \in (0,1)$ and $e^{\frac{x_u}{\lambda}} > 1$. These facts imply that $\mu_l > 0$. They also imply that $e^{\frac{x_u}{\lambda}} - 1 < e^{\frac{x_u}{\lambda}} - e^{\frac{-x_q}{\lambda}}$, and therefore that $\mu_h < 1$. Finally, $\mu_l = e^{\frac{-x_q}{\lambda}} \cdot \mu_h$, so $e^{\frac{-x_q}{\lambda}} < 1$ implies $\mu_l < \mu_h$.

Then, beginning with (1), the expressions for p_q^*, p_u^* are continuous on (μ_l, μ_h). Also, $\lim_{\mu \to \mu_l} p_q^*(\mu) = \lim_{\mu \to \mu_l} p_u^*(\mu) = 0$, with $p_q^* = p_u^* = 0$ for $\mu \leq \mu_l$, and $\lim_{\mu \to \mu_h} p_q^*(\mu) = \lim_{\mu \to \mu_h} p_u^*(\mu) = 0$, with $p_q^* = p_u^* = 1$ for $\mu \geq \mu_h$, so both p_q^* and p_u^* are continuous on $[0, 1]$. This implies that $p_q^* - p_u^*$ is also continuous on $[0, 1]$.

Then, for (2), we begin with $[0, \mu_l]$, on which $p_q^* = p_u^* = 0$, so $p_q^* - p_u^* = 0 \geq 0$. Then, for $[\mu_h, 1]$, $p_q^* = p_u^* = 1$, so $p_q^* - p_u^* = 0 \geq 0$ on this interval as well.

Next, we prove (3). On the interval (μ_l, μ_h), the relationship between p_q^* and p_u^* is derived above, with

$$p_u^* = \frac{e^{\frac{-x_q}{\lambda}} p_q^*}{e^{\frac{-x_q}{\lambda}} p_q^* + e^{\frac{x_u}{\lambda}} (1 - p_q^*)}.$$

This function has a positive first and second derivative in p_q^* on $[0, 1]$, so it is strictly convex, and it has fixed points at 0 and 1. This implies that $p_u^*(p_q^*)$ lies strictly under the line segment connecting $(0, 0)$ and $(1, 1)$. That is, $p_q^* - p_u^* > 0$ for $p_q^* \in (0, 1)$. Therefore, since $p_q^*(\mu) \in (0, 1)$ for $\mu \in (\mu_l, \mu_h)$, $p_q^* - p_u^* > 0$ on this interval.

Finally, for (4). We begin with the explicit expression for $p_q^* - p_u^*$ on (μ_l, μ_h):

$$p_q^* - p_u^* = \frac{1}{1 - e^{\frac{-x_q}{\lambda}}} \left(1 - \frac{e^{\frac{-x_q}{\lambda}} (e^{\frac{x_u}{\lambda}} - 1)}{\mu (e^{\frac{x_u}{\lambda}} - e^{\frac{-x_q}{\lambda}})} \right) - \frac{1}{1 - \mu} \left(\frac{\mu}{e^{\frac{x_u}{\lambda}} - 1} - \frac{e^{\frac{-x_q}{\lambda}}}{e^{\frac{x_u}{\lambda}} - e^{\frac{-x_q}{\lambda}}} \right).$$

Then, the second derivative of $p_q^* - p_u^*$ with respect to μ is

$$\frac{\partial^2}{\partial \mu^2} (p_q^* - p_u^*) = -2 \left[\frac{1}{\mu^3} \cdot \frac{e^{\frac{-x_q}{\lambda}} (e^{\frac{x_u}{\lambda}} - 1)}{(e^{\frac{x_u}{\lambda}} - e^{\frac{-x_q}{\lambda}})(1 - e^{\frac{-x_q}{\lambda}})} + \frac{1}{(1 - \mu)^3} \left(\frac{1}{e^{\frac{x_u}{\lambda}} - 1} - \frac{e^{\frac{-x_q}{\lambda}}}{e^{\frac{x_u}{\lambda}} - e^{\frac{-x_q}{\lambda}}} \right) \right].$$

$p_u^* > 0$ on (μ_l, μ_h) implies

$$\frac{\mu}{e^{\frac{x_u}{\lambda}} - 1} - \frac{e^{\frac{-x_q}{\lambda}}}{e^{\frac{x_u}{\lambda}} - e^{\frac{-x_q}{\lambda}}} > 0$$

$$\implies \frac{1}{e^{\frac{x_u}{\lambda}} - 1} - \frac{e^{\frac{-x_q}{\lambda}}}{e^{\frac{x_u}{\lambda}} - e^{\frac{-x_q}{\lambda}}} > 0$$

$$\implies \frac{\partial^2}{\partial \mu^2} (p_q^* - p_u^*) < 0, \ \mu \in (\mu_l, \mu_h).$$

Therefore, $p_q^* - p_u^*$ is strictly concave on (μ_l, μ_h). □

7.3 Proofs for Section 4

PROOF OF LEMMA 3:

First, formalize $(p_q^*(\mu, \gamma_c), p_u^*(\mu, \gamma_c))$ as the argmax of the function $R(p_q, p_u, \mu)$ subject to the constraint $\gamma_h(p_q, p_u, \mu) = \gamma_c$. This constraint is a continuous correspondence mapping μ and γ_c to possible values of p_q and p_u, and R is continuous in its arguments, so Berge's Maximum

Theorem implies that this argmax is upper-hemicontinuous in γ_c and μ. R is strictly concave for all $\mu \in (0,1)$, so this argmax is unique for any $\gamma_c \in (0, \mu_l]$, $\mu \in [0,1]$, which in turn implies that $(p_q^*(\mu,\gamma_c), p_u^*(\mu,\gamma_c))$ is strictly continuous for $\mu \in (0,1), \gamma_c \in [0,\mu_l]$ (with the assumption from Lemma (1) that $p_q^*(0) = 0, p_u^*(1) = 1$).

Furthermore, by the Implicit Function Theorem, for any μ, γ_c such that $(p_q^*(\mu,\gamma_c), p_u^*(\mu,\gamma_c)) \in (0,1)^2$, when the maximum is on the interior, $p_q^*(\mu,\gamma_c)$ and $p_u^*(\mu,\gamma_c)$ are differentiable in μ and γ_c. Since all optimal points satisfy Equation (14), this condition is equivalent to the condition that $p_q^*(\mu,\gamma_c) - p_u^*(\mu,\gamma_c) > 0$ when $\mu \in [0,1]$ and $\gamma_c \in (0,\mu_l)$.

It remains to show that there is an exactly one interval in $[0,1]$ on which $p_q^*(\mu,\gamma_c) - p_u^*(\mu,\gamma_c) > 0$ and that the function is concave in μ on that interval. Recall that the employer's problem is unchanged for $\mu \leq \gamma_c$ and $\mu \geq \mu_h$, so we need only consider (γ_c, μ_h). Note that on (γ_c, μ_l) the constraint is binding and changes equilibrium from $(p_q^*, p_u^*) = (0,0)$, so immediately to the right of γ_c we know that $p_u^*(\mu,\gamma_c)$ starts increasing in μ. By the Implicit Function Theorem we find that $\frac{\partial}{\partial\mu} p_u^*(\mu,\gamma_c) = \frac{(1-\gamma_c)}{(1-\mu)(\mu-\gamma-c)} \cdot p_u^*(\mu,\gamma_c) > 0$, so once the p_u^* starts increasing it continues to increase until $p_u^*(\mu_h', \gamma_c) = 1$ for some $\mu_h' \in (\gamma_c, \mu_h]$. If p_u^* were to decrease from 1 after hitting it, that would imply by the Mean Value Theorem that the derivative in μ is negative at some point in that interval, which cannot be as we have just seen. Therefore, (p_q^*, p_u^*) has an interior solution on the interval (γ_c, μ_h') for some μ_h'. Below this interval, $p_q^* = p_u^* = 0$, and above it $p_q^* = p_u^* = 1$.

Finally, $p_q^*(\mu,\gamma_c) - p_u^*(\mu,\gamma_c)$ is strictly concave on the interval (γ_c, μ_h'). Using Equation (14), we can rewrite $p_q^*(\mu,\gamma_c) - p_u^*(\mu,\gamma_c)$ as

$$p_q^*(\mu,\gamma_c) - p_u^*(\mu,\gamma_c) = \frac{(\mu-\gamma_c)}{(1-\gamma_c)\mu}(1 - p_u^*(\mu,\gamma_c)). \tag{20}$$

As stated above, $\frac{\partial}{\partial\mu} p_u^*(\mu,\gamma_c) = \frac{(1-\gamma_c)}{(1-\mu)(\mu-\gamma-c)} \cdot p_u^*(\mu,\gamma_c)$. Therefore, we can derive an explicit expression for $\frac{\partial^2}{\partial\mu^2}(p_q^*(\mu,\gamma_c) - p_u^*(\mu,\gamma_c))$:

$$\frac{\partial^2}{\partial\mu^2}(p_q^*(\mu,\gamma_c) - p_u^*(\mu,\gamma_c)) = -2\left(\frac{\gamma_c(1-p_u^*(\mu,\gamma_c))}{(1-\gamma_c)\mu^3} + \frac{\gamma_c \cdot p_u^*(\mu,\gamma_c)}{(\mu-\gamma_c)\mu^2(1-\mu)} + \frac{p_u^*(\mu,\gamma_c)}{\mu(1-\mu)^2}\right) \tag{21}$$

For any $0 < \gamma_c < \mu < 1$ and $p_u^* \in (0,1)$, this expression is strictly negative (as each term inside the parentheses is positive). Therefore, on the interval (γ_c, μ_h') where $p_q^*(\mu,\gamma_c) - p_u^*(\mu,\gamma_c) > 0$, the function is strictly concave. □

PROOF OF LEMMA 4:

Beginning with claim (1). Note that on the increasing side of $p_q^*(\mu,\gamma_c) - p_u^*(\mu,\gamma_c)$ for a fixed γ_c, all values of (p_q^*, p_u^*) are interior solutions satisfying Equation (14). Therefore, we can write

$\frac{\partial}{\partial \gamma_c}(p_q^*(\mu,\gamma_c) - p_u^*(\mu,\gamma_c))$ as

$$\begin{aligned}\frac{\partial}{\partial \gamma_c}(p_q^*(\mu,\gamma_c) - p_u^*(\mu,\gamma_c)) &= \frac{\partial}{\partial \gamma_c}\left(\frac{(\mu-\gamma_c)(1-p_u^*(\mu,\gamma_c))}{(1-\gamma_c)\mu}\right) \\ &= -\frac{(1-\mu)(1-p_u^*(\mu,\gamma_c))}{(1-\gamma_c)^2\mu} - \frac{(\mu-\gamma_c)}{(1-\gamma_c)\mu}\frac{\partial}{\partial \gamma_c}(p_u^*(\mu,\gamma_c))\end{aligned} \tag{22}$$

$\frac{\partial}{\partial \gamma_c}(p_u^*(\mu,\gamma_c))$ has an explicit expression by the Implicit Function Theorem, which is omitted here for space. Via simulation, I verified that for any parameters such that $\frac{\partial}{\partial \mu}(p_q^*(\mu,\gamma_c) - p_u^*(\mu,\gamma_c)) > 0$, it holds that $\frac{\partial}{\partial \gamma_c}(p_q^*(\mu,\gamma_c) - p_u^*(\mu,\gamma_c)) < 0$ (code available upon request).

Then, for claim (2). Since $p_q^*(\mu,\gamma_c) - p_u^*(\mu,\gamma_c)$ is concave on the interval on which it is positive-valued, there is a unique argmax μ^*. This μ^* satisfies the condition

$$\frac{\partial}{\partial \mu}(p_q^*(\mu,\gamma_c) - p_u^*(\mu,\gamma_c)) = \frac{\gamma_c}{(1-\gamma_c)\mu^2} - p_u^*(\mu,\gamma_c)\left(\frac{\gamma_c(1-\mu)+(1-\gamma_c)\mu}{(1-\gamma_c)\mu^2(1-\mu)}\right) = 0, \tag{23}$$

which implies that $p_u^* = \frac{\gamma_c(1-\mu^*)}{\gamma_c(1-\mu^*)+(1-\gamma_c)\mu^*}$. Since p_u^* still must satisfy Equation (17), we have the following equation implicitly defining μ^*:

$$\begin{aligned}&\gamma_c \log\left(\frac{\gamma_c^2(1-\mu^*)^2}{\gamma_c^2(1-\mu^*)^2 + (\mu^*-\gamma_c)(\gamma_c(1-\mu^*)+(1-\gamma_c)\mu^*)}\right) \\ &\quad + \log\left(1 + \frac{(\mu^*-\gamma_c)(\gamma_c(1-\mu^*)+(1-\gamma_c)\mu^*)}{\gamma_c(1-\mu^*)^2}\right) = \frac{x_u(1-\gamma_c) - x_q\gamma_c}{\lambda}.\end{aligned} \tag{24}$$

Also, by substituting $p_u^* = \frac{\gamma_c(1-\mu^*)}{\gamma_c(1-\mu^*)+(1-\gamma_c)\mu^*}$ into the equation $p_q^*(\mu,\gamma_c) - p_u^*(\mu,\gamma_c) = \frac{(\mu-\gamma_c)(1-p_u^*)}{(1-\gamma_c)\mu}$, we get the following expression for $m(\gamma_c)$:

$$m(\gamma_c) = p_q^*(\mu^*(\gamma_c),\gamma_c) - p_u^*(\mu^*(\gamma_c),\gamma_c) = \frac{\mu^*(\gamma_c) - \gamma_c}{\gamma_c(1-\mu^*(\gamma_c)) + (1-\gamma_c)\mu^*(\gamma_c)}. \tag{25}$$

The Implicit Function Theorem guarantees a derivative $\frac{\partial \mu^*}{\partial \gamma_c}$, and the Envelope Theorem guarantees that

$$\begin{aligned}\frac{\partial}{\partial \gamma_c} m(\gamma_c) &= \frac{\partial}{\partial \gamma_c}(p_q^*(\mu^*(\gamma_c),\gamma_c) - p_u^*(\mu^*(\gamma_c),\gamma_c)) \\ &= \frac{\frac{\partial \mu^*}{\partial \gamma_c} - 1}{\gamma_c(1-\mu^*) + (1-\gamma_c)\mu^*} - \frac{(\mu^*-\gamma_c)(1-2\mu^* + (1-2\gamma_c)\frac{\partial \mu^*}{\partial \gamma_c})}{(\gamma_c(1-\mu^*)+(1-\gamma_c)\mu^*)^2}\end{aligned} \tag{26}$$

By simulation, I find that for all valid $x_q, x_u, \lambda > 0, \gamma_c < \mu_l$, it holds that $\frac{\partial}{\partial \gamma_c} m(\gamma_c) < 0$ (again, code available upon request). Thus, as γ_c decreases, the maximum value attained by $p_q^*(\mu,\gamma_c) - p_u^*(\mu,\gamma_c)$ increases. □

PROOF OF PROPOSITION 1: Let M be as defined in Theorem (1), with $M = \max_{\mu \in [\mu_l,\mu_h]} (\mu^R(\mu,\mu_l) -$

μ), since $\gamma_c = \mu_l$ is equivalent to the unconstrained case. Since discrimination is feasible without intervention, we know that $M > 0$ and there exists a unique μ_{maxgap} with $\mu^R(\mu_{\text{maxgap}}, \mu_l) - \mu_{\text{maxgap}} = M$. Let $m^* = p_q^*(\mu_{\text{maxgap}}, \mu_l) - p_u^*(\mu_{\text{maxgap}}, \mu_l)$.

Let $m(\gamma_c)$ be as defined in Lemma (4). Since M is the interior maximum of $\mu^R(\mu, \mu_l) - \mu$, we know that $\frac{\partial}{\partial \mu}\mu^R(\mu, \mu_l) = 1$ at μ_{maxgap}. This implies that $m^* < m(\mu_l)$ and $\mu_{\text{maxgap}} < \bar{\mu}$ for $\bar{\mu}$ such that $p_q^*(\bar{\mu}, \mu_l) - p_u^*(\bar{\mu}, \mu_l) = m(\mu_l)$. By Lemma (4), the maximum $m(\gamma_c)$ is decreasing in γ_c, so for all $\gamma_c \in (0, \mu_l)$, $m(\gamma_c) \geq m(\mu_l)$. Therefore all values attained on the increasing side of $p_q^*(\mu, \mu_l) - p_u^*(\mu, \mu_l)$ are also attained precisely once on the increasing side of $p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c)$ for any $\gamma_c \in (0, \mu_l)$.

Thus, we can define $\mu^*(\gamma_c) = \min\{\mu \in (\gamma_c, \mu_h) : p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c) = m^*\}$. That is, $\mu^*(\gamma_c)$ solves the equation $G(\mu; \gamma_c) = p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c) - m^* = 0$. By, Lemma (3), we know that $p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c)$ is differentiable at μ^*. Therefore, the Implicit Function Theorem implies that $\mu^*(\gamma_c)$ is differentiable, and that

$$\frac{\partial \mu^*}{\partial \gamma_c} = -\frac{\frac{\partial G}{\partial \gamma_c}}{\frac{\partial G}{\partial \mu}} = -\frac{\frac{\partial}{\partial \gamma_c}(p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c))}{\frac{\partial}{\partial \mu}(p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c))}. \tag{27}$$

By definition, μ^* is on the side of $p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c)$ such that $\frac{\partial}{\partial \mu}(p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c)) > 0$. Then, by Lemma (4), it holds that $\frac{\partial}{\partial \gamma_c}(p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c)) < 0$ at μ^*. These facts then imply that

$$\frac{\partial \mu^*}{\partial \gamma_c} > 0. \tag{28}$$

So for any $\gamma_c < \mu_l$, it holds that $\mu^*(\gamma_c) < \mu^*(\mu_l) = \mu_{\text{maxgap}}$. Note that since $p_q^*(\mu^*(\gamma_c), \gamma_c) - p_u^*(\mu^*(\gamma_c), \gamma_c) = m^*$, it holds that $\mu^R(\mu^*(\gamma_c), \gamma_c) = \mu^R(\mu_{\text{maxgap}}, \mu_l)$ for all $\gamma_c \in (0, \mu_l)$.

This implies that for any γ_c,

$$\mu^R(\mu^*(\gamma_c), \gamma_c) - \mu^* \geq \mu^R(\mu_{\text{maxgap}}, \mu_l) - \mu_{\text{maxgap}} = M > 0.$$

Then, since $\mu^R(\gamma_c, \gamma_c) - \gamma_c = -\gamma_c < 0$ and $\mu^R(\mu_h, \gamma_c) - \mu_h = -\mu_h < 0$, the Intermediate Value Theorem implies that there are at least two points on (γ_c, μ_h) at which $\mu^R(\mu, \gamma_c) - \mu = 0$, multiple possible equilibria. Thus, for any γ_c, discrimination is always possible. □

PROOF OF PROPOSITION 2:

First, as in the proof of Lemma (3), formalize $(p_q^*(\mu, \gamma_c), p_u^*(\mu, \gamma_c))$ as the argmax of the function $R(p_q, p_u, \mu)$ subject to the constraint $\gamma_h(p_q, p_u, \mu) = \gamma_c$. This constraint is a continuous correspondence mapping μ and γ_c to possible values of p_q and p_u, and R is continuous in its arguments, so Berge's Maximum Theorem implies that this argmax is upper-hemicontinuous in γ_c and μ. R is strictly concave for all $\mu \in (0, 1)$, so this argmax is unique for any $\gamma_c \in [\mu_h, 1]$, $\mu \in (0, 1)$. At $\mu = 0$, $p_u^* = 0$, but p_q^* can be equal to any value. Similarly, at $\mu = 1$, $p_q^* = 1$, but p_u^* can be equal to any value. Since the constraint is not binding for any $\mu < \mu_l$, we are unconcerned

with such cases (so we ignore $\mu = 0$). For any $\gamma_c \neq 1$, the limiting p_u^* as μ goes to 1 is $p_u^* = 1$, so we can make (p_q^*, p_u^*) continuous on $\mu \in [\mu_l, 1], \gamma_c \in [\mu_h, 1)$ by setting $p_u^*(1, \gamma_c) = 1$.

As in Lemma (3), by a nearly identical argument, for a fixed γ_c we have that $p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c)$ is strictly positive and smooth on some interval (μ_l', γ_c) with $\mu_l' \in [\mu_l, \gamma_c)$. The function is equal to zero outside that interval. These proofs are left as an exercise to the reader.

Now, if we construct the realized qualification rate $\mu^R(\mu, \gamma_c) : [\mu_l, 1] \times [\mu_h, 1) \to [0, 1]$, with $\mu^R(\mu, \gamma_c) = F(w(p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c)))$, then μ^R is continuous on its domain. Then, $M(\gamma_c) = \max_{\mu \in [\mu_l, \gamma_c]} (\mu^R(\mu, \gamma_c) - \mu)$ is continuous in γ_c by Berge's Maximum Theorem for $\gamma_c \in [\mu_h, 1)$. Thus, if there exists any $\gamma_c \in (\mu_h, 1)$ such that $M(\gamma_c) < 0$, the Intermediate Value Theorem implies that there exists a $\gamma_c^* \in (\mu_h, 1)$ with $M(\gamma_c^*) = 0$. It remains to prove two claims: 1) $M(\gamma_c^*) = 0$ implies there is only one equilibrium under the affirmative action policy characterized by γ_c^*, and 2) $M(1) < 0$ implies that $M(\gamma_c) < 0$ for some $\gamma_c < 1$.

We begin with the first claim. Defining $d(\mu, \gamma_c) = p_q^*(\mu, \gamma_c) - p_u^*(\mu, \gamma_c)$, with what we have already established about $d(\mu, \gamma_c)$ it suffices to show that 1) $d(\mu, \gamma_c)$ is concave when it is positive valued, and 2) when $d(\mu, \gamma_c)$ is increasing, $\frac{\partial^3}{\partial \mu^3} d(\mu, \gamma_c) - 2\frac{(\frac{\partial^2}{\partial \mu^2} d(\mu, \gamma_c))^2}{\frac{\partial}{\partial \mu} d(\mu, \gamma_c)} < 0$, as argued in the proof for Theorem (1). In particular, these conditions guarantee that the second derivative of $\mu^R(\mu, \gamma_c)$ crosses zero at most once while $\mu^R(\mu, \gamma_c)$ is increasing in μ, and can only cross from positive to negative. I verified these properties by simulation (code available upon request). Therefore, when $M(\gamma_c) = 1$, $\mu^R(\mu, \gamma_c)$ has only one fixed point on $(0, 1)$.

Now we turn our attention to $M(1)$. With $\gamma_c = 1$, p_u^* *must* be equal to 0 for any $\mu \in (0, 1)$, so (p_q^*, p_u^*) approach $(1, 0)$ as μ goes to 1. For continuity of $p_u^*(\mu, 1)$, let $p_u^*(1, 1) = 0$. Then, we have

$$M(1) = \max_{\mu \in [\mu_l, 1]} (F(w \cdot p_q^*(\mu, 1)) - \mu) < 0.$$

We also have a closed form solution for $p_q^*(\mu, 1)$, obtained by simply differentiating the objective function and setting the first order condition equal to zero:

$$p_q^*(\mu, 1) = \frac{\mu - e^{\frac{-x_q}{\lambda}}}{\mu(1 - e^{\frac{-x_q}{\lambda}})},$$

for $\mu > e^{\frac{-x_q}{\lambda}}$, and $p_q^* = 0$ otherwise. Then, p_q^* is strictly increasing for $\mu \in (e^{\frac{-x_q}{\lambda}}, 1)$. By a similar condition on $\frac{\partial}{\partial \mu} p_q^*(\mu, 1)$, $\frac{\partial^2}{\partial \mu^2} p_q^*(\mu, 1)$, $\frac{\partial^3}{\partial \mu^3} p_q^*(\mu, 1)$ as seen above and in the proof of Theorem (1), we know that the second derivative of $\mu^R(\mu, 1)$ crosses zero at most once, and only from positive to negative. Our technical assumption guarantees a negative second derivative at $\mu = 1$, so we know that the first derivative is decreasing on some interval $(c, 1)$. Let $g_l = \mu^R(\frac{1+c}{2}, 1) - \mu^R(\frac{1+3c}{4})$ and $g_h = \mu^R(\frac{1+3c}{4}) - \mu^R(c, 1)$. Since the derivative of μ^R is greater at all points on $(c, \frac{1+c}{2})$ than at any point on $(\frac{1+c}{2}, 1)$, we know that $g_h > g_l$. Fix $\epsilon < \min\{\frac{g_h - g_l}{4}, |\frac{M(1)}{2}|\}$.

Now, for any $b \in (\mu_l, 1)$, we know that $\mu^R(\mu, \gamma_c)$ is continuous on $[\mu_l, b] \times [\mu_h, 1]$. Therefore, $\mu^R(\mu, \gamma_c)$ converges uniformly to $\mu^R(\mu, 1)$ as $\gamma_c \to 1$ on this compact interval for μ. Pick $b = 1 - \delta$,

where $\delta = \frac{1-c}{2} \cdot \min\{1, |\frac{M(1)}{4(g_l+2\epsilon)}|\}$. Then, by uniform convergence, let γ_c^* be such that for all $\gamma_c \in [\gamma_c^*, 1]$, $|\mu^R(\mu, 1) - \mu^R(\mu, \gamma_c)| < \epsilon$ for all $\mu \in [\mu_l, 1-\delta]$.

Then, the minimum difference between $\mu^R(\frac{1+3c}{4}, \gamma_c^*)$ and $\mu^R(c, \gamma_c^*)$ is $(\mu^R(\frac{1+3c}{4}, \gamma_c^*) - \epsilon) - (\mu^R(c, \gamma_c^*) + \epsilon) = g_h - 2\epsilon$. Similarly, the maximum difference between $\mu^R(\frac{1+c}{2}, \gamma_c^*)$ and $\mu^R(\frac{1+3c}{4}, \gamma_c^*)$ is $(\mu^R(\frac{1+c}{2}, \gamma_c^*) + \epsilon) - (\mu^R(\frac{1+3c}{4}, \gamma_c^*) - \epsilon) = g_l + 2\epsilon$. Since $\epsilon < \frac{g_h - g_l}{4}$, we know that $g_h - 2\epsilon > g_l + 2\epsilon$. Then, by the Mean Value Theorem, there exists a μ in $(c, \frac{1+3c}{4})$ with $\frac{\partial}{\partial \mu}\mu^R(\mu, \gamma_c^*) = \frac{4(g_h - 2\epsilon)}{1-c}$, and similarly, there exists a $\underline{\mu}$ in $(\frac{1+3c}{4}, \frac{1+c}{2})$ with $\frac{\partial}{\partial \mu}\mu^R(\mu, \gamma_c^*) = \frac{4(g_l+2\epsilon)}{1-c}$. Then, the Mean Value Theorem again implies that $\frac{\partial^2}{\partial \mu^2}\mu^R(\mu, \gamma_c^*) = \frac{4}{1-c}(g_l - g_h + 4\epsilon) < 0$ somewhere on the interval $(c, \underline{\mu})$. Since $\frac{\partial^2}{\partial \mu^2}\mu^R(\mu, \gamma_c^*)$ can change sign at most once, and only from positive to negative, we thus know that $\frac{\partial^2}{\partial \mu^2}\mu^R(\mu, \gamma_c^*) < 0$ for all $\mu \in (\underline{\mu}, \gamma_c^*)$.

Assume for contradiction that there exists some μ^* such that $\mu^R(\mu^*, \gamma^*) = \mu^*$. Since $\mu^R(\mu, \gamma_c^*) < \mu^R(\mu, 1) + \epsilon < \mu^R(\mu, 1) + |\frac{M(1)}{2}|$ for all $\mu \in [\mu_l, 1-\delta]$, and $\mu^R(\mu, 1) - \mu \leq M(1)$ for all μ, we know that μ^* must be on $(1-\delta, 1)$, where $\mu^R(\mu, \gamma^*)$ can deviate more than ϵ away from $\mu^R(\mu, 1)$. The *minimum* increase in $\mu^R(\mu, \gamma_c^*)$ required to reach the 45-degree line is $|\frac{M(1)}{2}|$, since at $\mu = 1-\delta$, $\mu^R(1-\delta, \gamma_c^*) - \mu \leq \frac{M(1)}{2}$. This increase must occur on an interval of length less than δ, since for all $\mu \geq \gamma_c^*$, $\mu^R(\mu, \gamma_c) = 0$. Thus, the Mean Value Theorem implies that at some $\bar{\mu} \in (1-\delta, \gamma_c^*)$, $\frac{\partial}{\partial \mu}\mu^R(\mu, \gamma_c^*) \geq \frac{|M|}{2\delta} > \frac{4(g_l+2\epsilon)}{1-c}$ by construction. This implies that there exists a $\mu \in (\underline{\mu}, \bar{\mu})$ with $\frac{\partial^2}{\partial \mu^2}\mu^R(\mu, \gamma_c^*) > 0$, which is a contradiction, as we established that $\frac{\partial^2}{\partial \mu^2}\mu^R(\mu, \gamma_c^*) < 0$ for all $\mu \in (\underline{\mu}, \gamma_c)$.

Therefore, we conclude that $\mu^R(\mu, \gamma_c^*) \neq \mu$ for any $\mu \in [\mu_l, 1]$. That is, $M(\gamma_c^*) < 0$. Then, finally, as stated above, the continuity of $M(\gamma_c)$ on $[\mu_h, 1)$ guarantees the existence of a γ_c' with $M(\gamma_c') = 0$ by the Intermediate Value Theorem, and discrimination is not possible in equilibrium when $\gamma_c = \gamma_c'$. □