

Meritocracy in the Face of Group Inequality*

Rajiv Sethi[†]

Rohini Somanathan[‡]

May 16, 2024

Abstract

Meritocratic systems are commonly understood as those that assign tasks to individuals who can best perform them. But future performance cannot be known prior to assignment, and must be inferred from other traits. We consider a model in which performance depends on two attributes—ability and training—where ability is endowed and unobserved and training is acquired and observed. The potential to acquire training depends on ability and resource access, so ability affects performance through two channels: indirectly through training and directly through the performance function. The population consists of two groups, each with the same ability distribution, but with differential access to resources. We show that performance-maximizing allocations are not necessarily group-blind or monotonic in training. This is true even when individuals can underreport scores or underinvest in training, and when commitment to selection policies is possible.

*We have benefitted from discussions with Kehinde Ajayi, Nageeb Ali, V. Bhaskar, Patricio Fernandez, Gerald Jaynes, Navin Kartik, Michael Lachmann, Glenn Loury, François Maniquet, Debraj Ray, John Roemer, Abhiroop Sarkar, Thomas Scanlon, and Sahana Subramanian. For fellowship support, Sethi thanks the Radcliffe Institute for Advanced Study at Harvard University, and Somanathan thanks the Center for Advanced Study in the Behavioral Sciences at Stanford University.

[†]Department of Economics, Barnard College, Columbia University and the Santa Fe Institute.

[‡]Department of Economics, Delhi School of Economics and 2022-23 Annenberg Fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford University.

1 Introduction

Policy debates in unequal societies are often about acceptable trade-offs between rewarding merit and encouraging the entry of resource-disadvantaged groups into privileged positions. Meritocratic systems are commonly understood as those that provide jobs and opportunities based on talent without regard to indicators such as race and sex. It has been argued that such systems are non-discriminatory because they allow for fair competition among eligible pools of candidates. Based on the non-discrimination argument, several recent court rulings have dismantled college admissions systems which conditioned acceptance probabilities on group-membership.

The above conception of a meritocracy ignores two important aspects of most practical assignment problems. First, observable measures of performance are based on past achievement and are noisy signals of future performance. Since selection takes place *before* tasks are performed, efficiency must be defined in terms of *expected* performance maximization and decisions about candidates made with imperfect information about their relevant attributes. Information about group identity can help solve this inference problem and lead to better predictions of future productivity.

Second, the usefulness of conditioning on group-membership depends on the goals of the organization and the tasks that need to be performed. Steven Durlauf emphasizes that “merit needs to be assessed relative to the properties of the human capital production functions” and Thomas Scanlon expresses a similar idea from a philosopher’s perspective: “What counts as a talent in the relevant sense will depend not only on the goals of the institution but also on the way in which this institution and the particular position in question are organized.”¹

We characterize meritocratic allocations in a setting with two social groups and limited information about the future performance of applicants. We model performance as a function of *ability*, which is unobserved and *training* which can be observed. Applicants in each group have the same distribution of ability, but differ in their training distributions because these depend additionally on their access to resources such as those embodied in school and neighborhood environments, and resource access varies by group. We are interested in the conditions under which selection rules that maximize expected performance are group-contingent. We show that group-blind selection is efficient only in special cases, and that the representation of resource disadvantaged groups depends on

¹See Durlauf (2008, p. 138) and Scanlon (2018, p. 45) respectively.

the combined effect of the performance (or the human capital production) function, the share of total applicants selected, and the asymmetry in resource access across groups.

A critical assumption of our model is that some determinants of performance are observed, while others must be inferred. Able candidates who have been deprived of the types of resources that shape measurable metrics of achievement might exhibit large gaps between their past training and their future performance. Mindful of this, selectors may want to skip over some training levels to reach them. When social groups have differential access to resources, this logic leads to group-contingent selection policies. We show that these effects persist, and can even be amplified, if candidates respond to non-monotonic selection policies by strategically underinvesting in training.

We see this paper as making two main contributions. The first is to clarify the concept of meritocracy. As long as access to resources varies within a population and performance depends both on resources and non-resource based endowments, we show that efficient allocations are *not* generally monotonic in training. This holds irrespective of whether there are identifiable social groups in the population. We further show that in the presence of group inequality, this logic can result in group-contingent selection. Departures from monotonicity and group-blindness do not necessarily reflect a neglect of meritocratic ideals. Moreover, the degree of representation of disadvantaged groups is not a summary statistic for the extent to which efficiency has been foregone in favor of greater equality.

Our work is related to an extensive literature on optimal selection rules in the face of group inequality. In the next section, we discuss some connections between this literature and our paper. Section 3 lays out our basic model and Section 4 derives the solution to the assignment problem when individual training choices are not strategic. In Section 5, we allow agents to strategically underinvest in training and derive equilibria of the signaling game. We conclude in Section 6 with a discussion of open questions for future research in this area.

It is worth emphasizing that we do not propose meritocratic assignment in place of other more equitable allocation rules. Rather, we seek to characterize the nature of efficient allocations since these are critical benchmarks in the larger debate on social equality. Our paper is motivated by current disagreements related to the meaning of meritocracy in academic and popular discourse and by the laws and institutions that govern the implementation of meritocratic principles in unequal societies.

2 Related Literature

The concept of merit has been extensively discussed and variously interpreted. Richard Arneson (2015) uses the term “formal equality of opportunity” to refer to selection of the most qualified for any given position. John Roemer refers to assignment based on talent as the non-discrimination principle, according to which “race or sex, as such, should not count for or against a person’s eligibility for a position, when race or sex is an irrelevant attribute insofar as the performance of the duties of the position is concerned” (Roemer, 1998, p.1). We use this definition as a starting point and illustrate why group identity may be relevant for selection for informational rather than intrinsic reasons.

Steven Durlauf, Thomas Scanlon and Amartya Sen provide theoretical and philosophical foundations for a context-dependent and forward-looking definition of merit. In evaluating the efficiency implications of alternative assignment rules, Durlauf argues for a “move from merit as reward to merit as effectiveness” and argues that race-conscious admissions can improve efficiency (Durlauf, 2008, p. 139).² Scanlon (2018) emphasizes that the notion of talent relevant for procedural fairness depends on the nature of the positions for which individuals are being selected. Sen (2000) argues that merit cannot be defined independently of social goals and that the term should be assigned to actions that promote these goals. The origins of these debates go back at least as far as Aristotle, who said: “For it is to the function that the superiority should contribute, and superiority in wealth and in good-breeding contribute nothing to it.”³

In the area of optimal selection rules, there is a sizable literature on identity-contingent admissions and hiring under conditions of imperfect information about potential candidates. Early work focused on statistical discrimination (Phelps, 1972; Arrow, 1973; Coate and Loury, 1993; Aigner and Cain, 1977; Cornell and Welch, 1996). More recently, there has been a focus on optimal selection rules when diversity and merit are both valued (Chan and Eyster, 2003; Fryer et al., 2008; Fryer and Loury, 2013). These papers ask how traditional or “sighted” affirmative action compares with color-blind affirmative action, which refers to policies that are not explicitly group-contingent, but are nevertheless motivated by diversity goals. In this literature, individual merit is treated as synonymous with some observable qualification such as a test score. For instance, Chan and Eyster (2003,

²Cavanagh (2002) discusses alternative notions of merit at some length and argues that scholars using the term often confuse a conception of merit related to moral desert with one that relies on the value of individual characteristics in relation a particular occupation.

³*Politics* 111.12, 1283a1-3. Translation by Reeve (2017).

p. 860) assume that the “expected academic promise... of a candidate with test score t is simply t : the higher a candidate’s score, the higher her quality.” In this framework, meritocratic allocations necessitate the application of a common qualification threshold to all members of the population, regardless of identity.

Cestau et al. (2017) study selection decisions made by an (unnamed) school district. They distinguish between *profiling*, which is an attempt to use demographic information to better meet performance goals unrelated to diversity, and *affirmative action*, which involves preferential treatment for a group beyond levels justified by profiling. The former serves the goal of maximized performance, while the latter increases diversity relative to performance-maximizing levels. They find that the district engages in both profiling and affirmative action with respect to family income, and engages in profiling but not affirmative action with respect to race. One of our main points is that bans on the use of group membership in the process of selection eliminate both affirmative action and performance profiling. Contrary to the rhetoric surrounding such initiatives, such mandates potentially block the implementation of meritocratic allocations.

A number of popular writings have highlighted the extent to which test scores reflect household and school resources rather than aptitude (Sandel, 2020; Markovits, 2019). Also related are studies by social psychologists that document the extent to which grades and test scores underestimate the potential of disadvantaged populations to perform. They argue that resources available to students in academic settings are both material and psychological and illustrate ways in which building supportive psychological environments, a process that has been termed “affirmative meritocracy”, can bridge the gap between measured academic grades and potential performance (Aronson and Steele, 2005; Walton et al., 2013). These writings support our assumption that the observable correlates of productivity are noisy measures of actual productivity.

To summarize, we define a meritocratic allocation as one that selects the most capable individuals from a pool of candidates. Like much of the literature on optimal selection, we assume performance is based on individual attributes. We depart from this literature in assuming that only some of these attributes can be observed and explore the role of group identity in maximizing expected performance.

3 The Model

Consider a population composed of two groups, 1 and 2, with shares s_1 and s_2 respectively. Individuals within each group differ in their ability a , and their past access to resources r . Both groups have the same distribution of ability, with distribution function $F(a)$. The distribution of resources varies by group. Most of our analysis is based on two resource levels, r_l and r_h where $r_l < r_h$. The proportion of individuals in group i with access to the higher resource level is denoted by q_i . We assume $q_1 < q_2$ and refer to the first group as *disadvantaged*.⁴

Ability and resources are independently distributed and neither can be directly observed. We observe only group identity and a signal t , which we refer to as *training*. This could be thought of as a measure of educational attainment, a test score, or any combination of productivity-relevant metrics. The highest attainable level of training for an individual is given by the continuous function $\tau(a, r)$, which is increasing in both arguments. That is, at each level of ability and resources, the level of training satisfies

$$t \leq \tau(a, r). \quad (1)$$

We are interested in the problem of selecting a fraction k of the applicant pool into scarce positions, and refer to k as *elite capacity*. Performance in these positions is increasing in both ability and training and given by the function

$$p = \phi(a, t). \quad (2)$$

Ability therefore has both a direct and an indirect effect on performance, capturing the idea that past scores might underrepresent the capacity of talented individuals to perform.

Any selection policy can be described by a pair of functions $\pi_i(t)$ that denote the probability of being selected conditional on exhibiting training t and belonging to group $i \in \{1, 2\}$. The selection policy that maximizes expected performance depends on the training and performance functions, $\tau(a, r)$ and $\phi(a, t)$, and on whether or not candidates choose training strategically. We refer to a selection policy as *monotonic* if, for each group i , there is a threshold training level which separates selected and rejected candidates.⁵ A

⁴The assumption of two resource levels simplifies our analysis and exposition. We consider continuously varying resource levels in the Appendix.

⁵Selection at the threshold might be random if the distribution of training is discrete, which can happen

policy is *group-blind* if the likelihood of selection is independent of group membership: $\pi_1(t) = \pi_2(t)$ at all t .

If a policy is not group-blind, we say that it is *group-sighted*. A group-sighted policy *favors the disadvantaged group* if $\pi_1(t) \geq \pi_2(t)$ at all t , with strict inequality at some t . In the next section we show that policies that maximize expected performance might be neither monotonic nor group-blind, and will favor the disadvantaged group under certain conditions.

Since the performance function is increasing in training, it is worth considering how a performance-maximizing selection policy could be non-monotonic in training. The answer lies in the inference about ability drawn at any given training level. If training is resource-intensive, while performance relies heavily on ability, an optimal selection policy might ignore some candidates with high training levels in favor of others with somewhat lower training.

A non-monotonic selection policy of this kind, however, can create incentives for underinvestment in training. Unless the consumption benefits of attaining one's potential are sufficiently great, those with low ability and high resources may underinvest in training to pool with those below them to increase their likelihood of selection. This in turn affects the equilibrium assignment rule. We explore strategic underinvestment in Section 5, but first characterize optimal selection in the absence of strategic training choices.

4 Selection and Representation

4.1 Optimal Selection Policies

In this section, we assume that the ability distribution (common to both groups) is continuous with strictly positive density $f(a)$ and support $[0, 1]$. We also assume that all applicants exhibit the maximal training achievable with their ability and resource access. This results in deterministic optimal selection rules; each candidate is selected with probability zero or one.⁶

if both ability and resources have discrete distributions, or if training is chosen strategically. See Section 5 for analysis of these cases.

⁶Optimal selection is non-deterministic in Chan and Eyster (2003) and Fryer and Loury (2013) despite continuous (and exogenously given) score distributions, because the authors assume that only monotonic

In the absence of strategic underinvestment, the level of training for each applicant is uniquely determined as $t = \tau(a, r)$. For each t , we define the ability levels $\alpha_l(t)$ and $\alpha_h(t)$ that allow an applicant to achieve training t with low and high resources respectively:

$$t = \tau(\alpha_l(t), r_l) = \tau(\alpha_h(t), r_h).$$

Clearly $\alpha_l(t) > \alpha_h(t)$ and both α_h and α_l are strictly increasing in t ; given resource access, higher ability is reflected in higher training.

Meritocratic selection requires choosing applicants with the highest expected performance. In any given group, those with $a = 1$ and high resources achieve the highest possible level of training, denoted $\bar{t} = \tau(1, r_h)$. Those with $a = 1$ and low resources achieve $t^* = \tau(1, r_l) < \bar{t}$. As a result, those with training $t \in (t^*, \bar{t}]$ reveal themselves to have high resource access, with performance $\phi(\alpha_h(t), t)$. At training levels $t < t^*$, there is a mix of applicants with low and high resources. The performance of those with low resources is given by $\phi(\alpha_l(t), t)$ which is higher than the performance of those with high resources, $\phi(\alpha_h(t), t)$, since $\alpha_l(t) > \alpha_h(t)$. These differences in applicant composition above and below $t = t^*$ result in a drop in expected performance at this training level. The productivity gains from training slightly higher than t^* are swamped by the fall in expected ability of the applicant pool. Hence there exists a training interval above t^* within which performance is lower than expected performance at t^* . We now identify this interval, which plays a key role in our results.

Let $\gamma_i(t)$ denote the likelihood that an individual in group i , with training t , has high resource access. The expected productivity of someone in group i with training $t < t^*$ is then

$$E(p_i|t) = \gamma_i(t)\phi(\alpha_h(t), t) + (1 - \gamma_i(t))\phi(\alpha_l(t), t). \quad (3)$$

We obtain a closed form expression for $\gamma_i(t)$ in the appendix, and show that at any $t \leq t^*$, $\gamma_1(t) < \gamma_2(t)$. This, together with (3), implies

$$E(p_1|t) > E(p_2|t). \quad (4)$$

That is, at any training level at which both high and low resource types are pooled, the disadvantaged group is less likely to have high resource access, and therefore has higher expected performance than the advantaged group.⁷ This is intuitive—those with lower resources need higher ability to reach any given training level t , and since ability and

policies are feasible. Relaxing this constraint results in generically deterministic selection in the absence of strategic underinvestment (Ray and Sethi, 2010).

⁷This is stated and proved as Lemma 1 in the appendix.

resources are independently distributed, the disadvantaged group has a greater share of higher ability members at any given training level.

Before formally characterizing the optimal selection policy, we highlight some of its features using the following example.

Example 1. Suppose that the two groups are of equal size and ability is uniformly distributed on $[0, 1]$. The training and performance functions are linear, with $t = ar$, and $p = \beta a + (1 - \beta)t$, with $\beta = 4/5$. The two resource levels are $r_l = 1$ and $r_h = 3/2$ and the shares with high resources in the two groups are $q_1 = 1/5$ and $q_2 = 2/3$. Then $t^* = 1$, $\bar{t} = 3/2$, $\alpha_h(t) = 2t/3$, and $\alpha_l(t) = t$. The performance for these two levels of ability is given by $\phi(\alpha_h(t), t) = (11/15)t$ and $\phi(\alpha_l(t), t) = t$ respectively. For $t \leq t^*$, $\gamma_1(t) = 1/7$ and $\gamma_2(t) = 4/7$, and hence $E(p_1|t) = 101t/105$ and $E(p_2|t) = 89t/105$. Since the weights γ_i do not depend on t , the expected performance functions are linear.⁸

Figure 1 shows how expected performance varies with training for the parameters in the above example. Blue and red lines represent the expected performance of groups 1 and 2 respectively for training levels below t^* . Since all training levels above t^* can only be achieved with high resources, candidates from both groups have the same performance in this range, and this is given by the yellow line. We see that the disadvantaged group has higher expected performance for all training levels below t^* .

We now define two critical training levels (one for each group) that play an important role in our analysis. Given our assumptions, expected performance varies continuously over $t \in [0, t^*]$, then drops discontinuously at t^* due to the change in the composition of the applicants with respect to resource-access, and then increases continuously again until the highest attainable score, $\bar{t}$. Hence there is a unique $\hat{t}_i \in (t^*, \bar{t})$ such that

$$\phi(\alpha_h(\hat{t}_i), \hat{t}_i) = \max_{t \leq t^*} E(p_i|t). \quad (5)$$

When expected performance is strictly increasing in training for all $t < t^*$, these training thresholds are easily visualized, as shown in Figure 1. However, even when the expected performance functions are non-monotonic, (4) implies that they are unambiguously ordered:

$$t^* < \hat{t}_2 < \hat{t}_1 \quad (6)$$

⁸The values of $\gamma_i(t)$ in the example may be obtained using (13) in the appendix, which simplifies in the case of uniform ability to

$$\gamma_i(t) = \frac{q_i \alpha'_h(t)}{q_1 \alpha'_h(t) + (1 - q_1) \alpha'_l(t)}$$

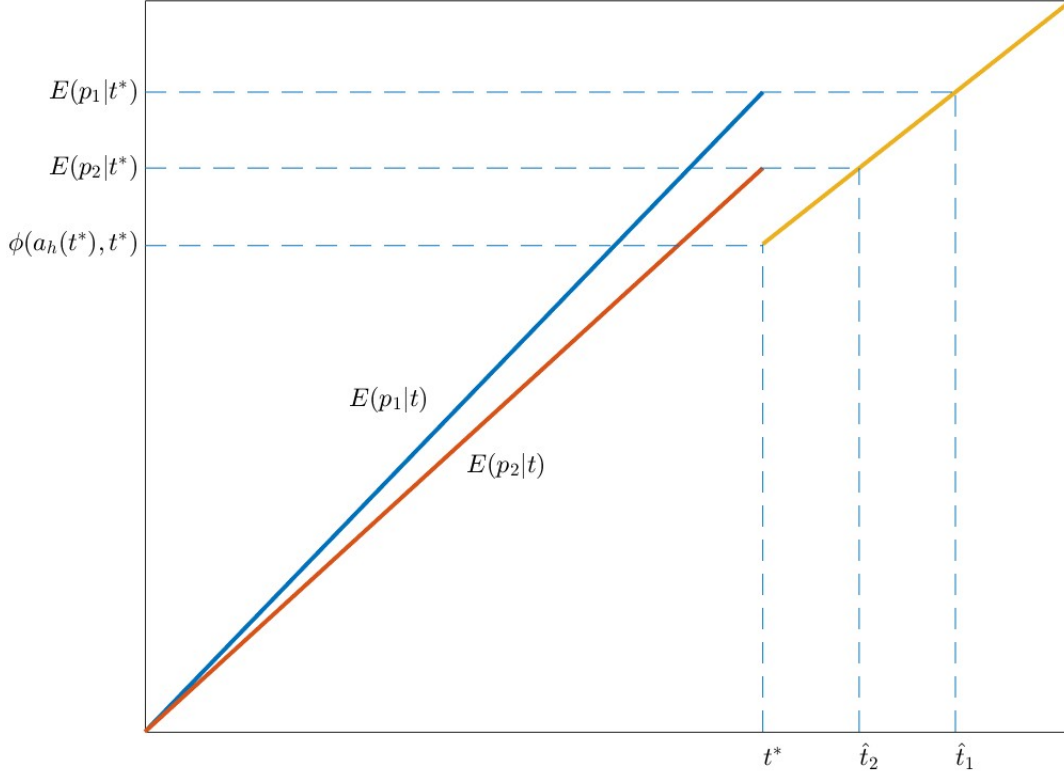


Figure 1: Expected Performance Conditional on Training for Two Groups

We now use these critical training levels to define a threshold level of elite capacity that is used to characterize the optimal selection policy. Let $\hat{k}$ denote the proportion of the applicant population (aggregated across both groups) that has training level at least $\hat{t}_1$. That is,

$$\hat{k} = \sum_{i=1}^2 s_i q_i (1 - F(\alpha_h(\hat{t}_1))). \quad (7)$$

If elite capacity $k \leq \hat{k}$, then all available positions can be filled by candidates with training $t \geq \hat{t}_1$. In this case (but not otherwise) the performance-maximizing policy is group-blind:

Proposition 1. *The performance-maximizing policy is monotonic in training and group-blind if elite capacity $k < \hat{k}$, and group-sighted and favors the disadvantaged group if $k > \hat{k}$.*

Figure 2 illustrates Proposition 1 by tracing the sets of training levels of selected applicants that accompany an expansion of elite capacity. Dashed lines mark the lowest expected performance among those selected. Intervals which include both groups are marked in yellow on the horizontal axis and those with only the disadvantaged group

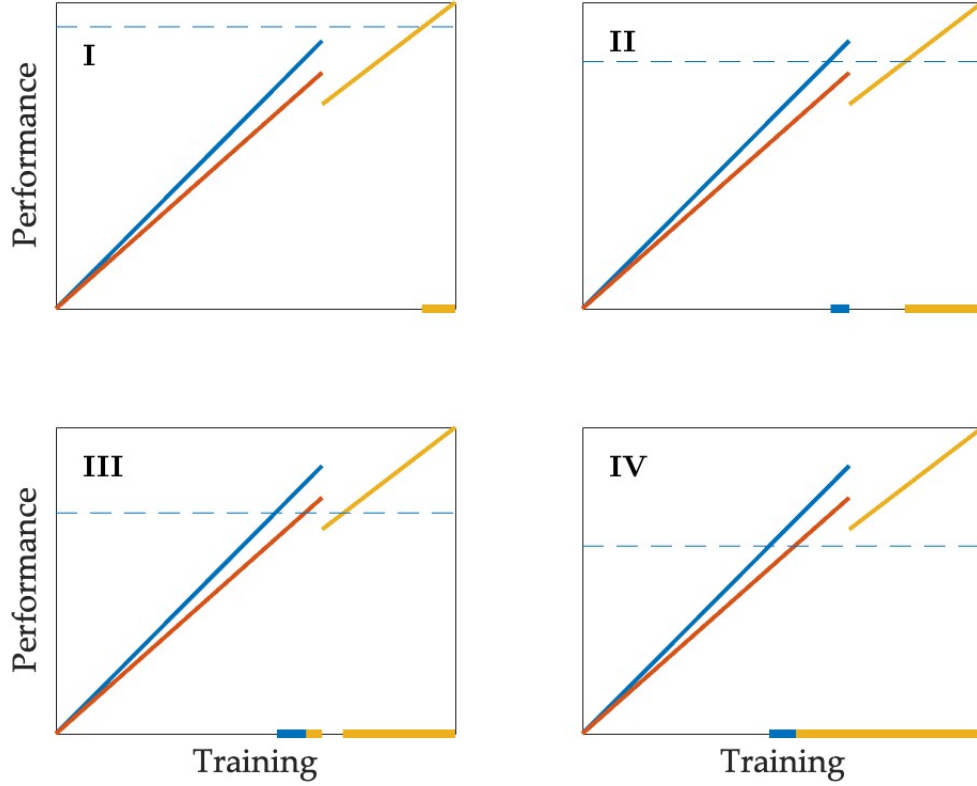


Figure 2: Optimal Selection as Elite Capacity Varies

are blue. As implied by (4), there are no intervals over which only the advantaged group is selected.

The performance-maximizing allocation lies in one of four regimes. The first quadrant has $k < \hat{k}$ so selection is monotonic and group blind. For elite capacity slightly larger than $\hat{k}$, selection for the advantaged group is monotonic, but the disadvantaged group faces non-monotonic selection. This is shown in the second quadrant. The third quadrant shows both groups facing non-monotonic selection with a further expansion in elite capacity. Notice that the standard applied to the disadvantaged group is less restrictive: some individuals with low levels of training are selected only if they belong to the disadvantaged group. Finally, the last quadrant shows that when elite capacity is sufficiently large, the performance-maximizing allocation is again monotonic. But it is not group-blind: the threshold for selection from the disadvantaged group is lower than for the advantaged group.

4.2 Representation

It is clear that the disadvantaged group will be underrepresented under any monotonic and group-blind policy, and hence underrepresented when elite capacity is smaller than $\hat{k}$. However, as seen in the previous section, further expansions in capacity could favor this group via a selection effect which mitigates their resource disadvantage. The size of the selection effect depends on the importance of ability relative to training in determining performance.

Define the *relative representation* ρ_i of group i as the fraction of individuals from this group who are selected, divided by the elite capacity k . Then, by definition,

$$s_1\rho_1 + s_2\rho_2 = 1.$$

Proportional representation or statistical mirroring arises when $\rho_1 = \rho_2 = 1$. In this case a fraction k from each group is selected, and those in elite positions reflect the demographic composition of the population at large. The disadvantaged group is underrepresented (relative to its share of the population) if $\rho_1 < 1 < \rho_2$.

Our measure of relative representation ρ_i may be interpreted as follows: it is the measure of those selected from group i relative to the measure that would be selected under statistical mirroring. This value ranges from zero (if no members are selected) to $1/s_i$ (if all members of group i and no members of the other group are selected). This can be very large if a small group occupies all elite positions.

To illustrate, suppose that elite capacity $k < \hat{k}$. In this case selection is monotonic and group-blind, and there is some training threshold $t > \hat{t}_1$ above which all are selected and below which all are rejected. Furthermore, only those with high resource access can reach this threshold. In this case,

$$\rho_i = \frac{q_i(1 - F(\alpha_h(t)))}{k} = \frac{q_i(1 - F(\alpha_h(t)))}{(s_1q_1 + s_2q_2)(1 - F(\alpha_h(t)))} = \frac{q_i}{s_1q_1 + s_2q_2}. \quad (8)$$

Note that this is independent of t (and hence of k). That is, when elite capacity is below $\hat{k}$, the relative representation of each group is constant. When the two groups are of equal size, this simplifies to $\rho_i = 2q_i/(q_1 + q_2)$, or twice the share of group i among high resource individuals in the population at large. For the specification in Example 1, the share of the disadvantaged group among high resource individuals is $3/13$, and hence the relative representation of this group when k is sufficiently small is $6/13 \approx 0.46$. At the other extreme, as elite capacity becomes very large, the relative representation of both groups approaches 1, since almost all applicants are selected.

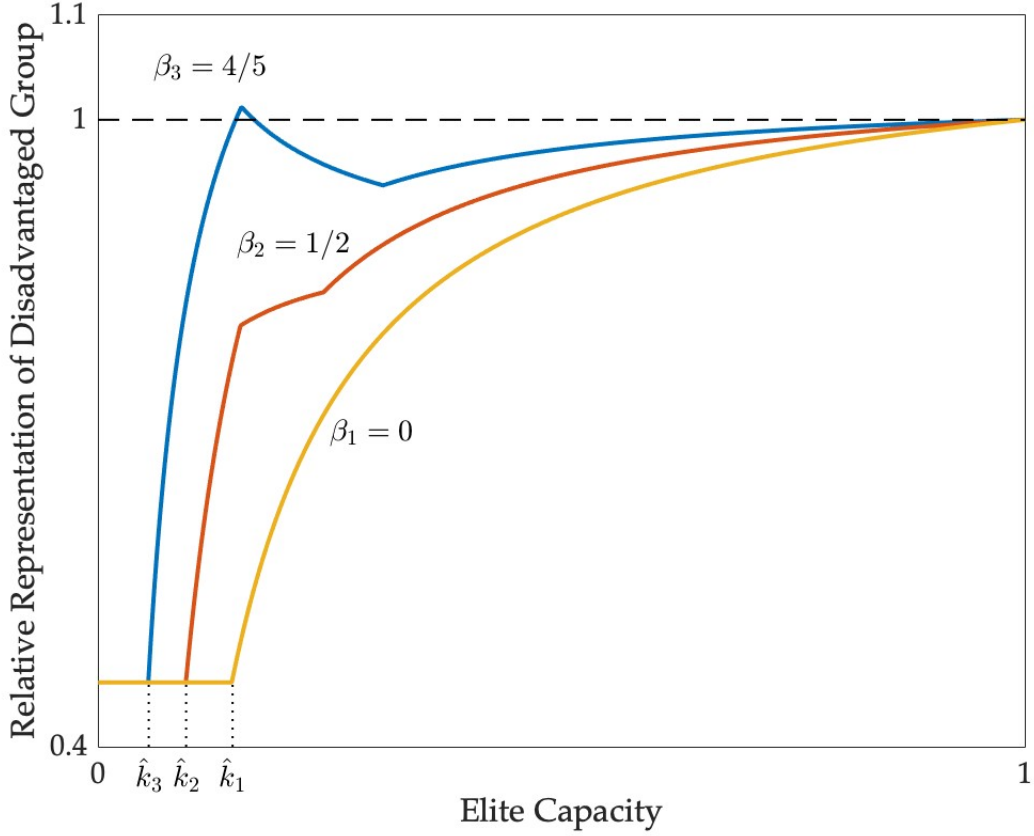


Figure 3: Relative representation of the disadvantaged group for different values of β .

Between these two extremes, relative representation varies with elite capacity in interesting and complex ways. We explore this using the parametric specification of the performance function from Example 1, in which β reflects the relative importance of ability:

$$p = \phi(a, t) = \beta a + (1 - \beta)t.$$

The following result states that higher values of β restrict the range of capacity levels over which allocations are necessarily monotonic and group blind:

Proposition 2. *If the performance function is given by $\phi(a, t) = \beta a + (1 - \beta)t$, the threshold level of elite capacity $\hat{k}$ below which the performance-maximizing allocation is monotonic and group-blind is decreasing in β .*

For $k > \hat{k}$, relative representation depends on elite capacity. Figure 3 uses the same training and ability distributions as in Example 1 and traces the relationship between relative representation of the resource disadvantaged group and elite capacity for three

alternative values of β . The horizontal dashed line in the figure corresponds to proportional representation, where the same proportion k from each group is selected. The lowest curve corresponds to $\beta = 0$, so only training matters for productivity. This results in performance maximizing policies being both monotonic and group-blind, with the disadvantaged group always underrepresented. When ability also has an independent effect on productivity the disadvantaged group benefits from the screening effect, to a degree that is increasing in β . For high enough values of β , there is a range of capacity levels for which the meritocratic allocation results in *overrepresentation* of the disadvantaged group. That is, the screening effect favoring a disadvantaged group can be strong enough to overcome the disadvantage itself.

The figure also illustrates how representation of the disadvantaged group varies with the regimes identified in Figure 2. When β is sufficiently high, transitions from the second regime (in which only the resource disadvantaged group faces non-monotonic selection) to the third (in which both groups do), reduce the relative representation of the disadvantaged group. Eventually, as k approaches 1, all applicants secure positions and we have proportional representation.

5 Strategic Underinvestment

The previous section assumes that all candidates reach the highest level of training that their ability and resources permit. However, when the selection policy is not monotonic, there may be a potential gain from underinvesting in training so selection and training choices are jointly determined in equilibrium. In this section, we characterize the equilibria of this game.

To simplify our analysis, we limit ourselves to two levels of ability a_l and a_h , where $a_l < a_h$. In each group, the proportion of individuals with low ability is $F(a_l)$. Let $t_{ij} = \tau(a_i, r_j)$ denote the maximum attainable training level for someone with ability a_i and resources r_j . Since we have two resource levels, there are four possible values of t_{ij} , which we can rank as follows:

$$t_{ll} < t_{lh}, \quad t_{hl} < t_{hh}.$$

Note that the two intermediate values of t_{ij} cannot be unambiguously ranked: those with low ability and high resources may be able to attain higher levels of training than those with high ability and low resources. This will happen if training is heavily resource de-

pendent and resource differences are large. Suppose that this is the case, and we have

$$t_{hl} < t_{lh}. \quad (9)$$

If performance depends heavily on ability, and less so on training, then it is possible that high ability types at training level t_{hl} may be preferred to low ability types at the higher training level t_{lh} . Suppose that this too is the case, and we have

$$\phi(a_h, t_{hl}) > \phi(a_l, t_{lh}). \quad (10)$$

When (9) and (10) hold, and there is no strategic underinvestment, selection for performance could involve skipping over those at training t_{lh} to access the higher ability types at training t_{hl} as the following example illustrates.

Example 2. Suppose that the two groups are of equal size and ability levels are $a_l = 3/4$ and $a_h = 1$, with $F(a_l) = 0.7$. The two resource levels are $r_l = 1$ and $r_h = 3/2$ and the shares with high resources in the two groups are $q_1 = 1/10$ and $q_2 = 3/10$. Elite capacity is $k = 0.1$. The training and performance functions are those of Example 1: $t = ar$, and $p = \beta a + (1 - \beta)t$, with $\beta = 4/5$. In the absence of strategic underinvestment, it is easy to show that the performance-maximizing selection policy will set $\pi_i(t_{hh}) = 1$ and $\pi_i(t_{ll}) = \pi_i(t_{lh}) = 0$ for both groups, and $\pi_i(t_{hl}) > 0$ for one or both groups.

Notice that (9) and (10) are both satisfied in this example since

$$t_{hl} = 1 < 9/8 = t_{lh} \quad \text{and} \quad \phi(a_h, t_{hl}) = 1 > 33/40 = \phi(a_l, t_{lh}).$$

With elite capacity at 10%, the 6% with training t_{hh} will all be selected. The remaining 4% will be drawn from those having training t_{hl} , which lies below t_{lh} . This draw could (but need not) be uniform at random without regard to group membership, in which case the policy would be group-blind. But regardless of how the remaining 4% are selected, the performance-maximizing policy will not be monotonic. It will involve accepting all those at the highest level of training, along with some of those at the third highest level, skipping over the individuals with low ability and high resources who are at the second highest level of training.

The non-monotonic policy described above might induce the individuals who have been skipped over to underinvest, and pool with those who are selected with positive probability at lower training levels. We now consider equilibrium policies in the presence of such strategic investments in training.

Note that underinvestment lowers the expected performance of the pool at t_{hl} , which now contains both high ability low resource and low ability high resource types. It does so for two reasons: first, because individuals who were initially at training t_{lh} had lower performance than those with whom they are pooling, and second, because they are lowering their performance even further by underinvesting in training relative to their potential. As long as this lowered pool quality at t_{hl} remains above the performance of those who were initially at t_{lh} , however, there can be an equilibrium with such pooling. The exact conditions for this are identified below, but for the purposes of the example, suppose that these conditions are satisfied in both groups. Then, for the parameter values given in Example 2, there is an equilibrium in which all individuals with low ability and high resources choose training level t_{hl} rather than t_{lh} , thus pooling with those at a training level below their potential. Despite being diluted, this pool has higher expected performance (in both groups) than those *known* to have low ability and high resources.

The strategic training investments described above also create differences in expected performance across groups. Performance at t_{hl} is higher for the disadvantaged group because it has a larger share of candidates with high ability and low resources. The optimal policy therefore favors this group, choosing from the advantaged group only once the supply of candidates from the disadvantaged group has been exhausted. In fact, this effect can be strong enough for the disadvantaged group to be overrepresented for certain values of elite capacity, despite having a worse distribution of training, and an identical distribution of ability. And this effect arises even though the only criterion for selection is the maximization of expected performance.

If commitment to a selection policy were possible, the inefficiency associated with underinvestment can be avoided by applying the same probability of selection to the two intermediate training levels in any given group. All candidates would now achieve their maximum level of training. The composition of the selected pool would be the same as in the case without commitment, so the disadvantaged group is still preferred at these two training levels.⁹

We now proceed to a fuller characterization of possible equilibria. In order to do this, we define two thresholds, similar to those in Section 4. We denote by $\bar{k}$ the proportion of candidates having both high ability and high resources:

$$\bar{k} = (1 - F(a_l))(s_1 q_1 + s_2 q_2).$$

⁹See Sethi and Somanathan (2023) for a detailed discussion and visualization of this example for all three cases—no underinvestment, pooling, and commitment.

This is the largest level of elite capacity that can be filled with those with the highest level of training. Our second threshold, denoted by $\tilde{k}$, is the proportion of candidates with either high ability or high resources (or both):

$$\tilde{k} = 1 - F(a_l)(s_1(1 - q_1) + s_2(1 - q_2)).$$

When $k < \bar{k}$, elite capacity is smaller than the number of candidates with the highest level of training. Since these individuals also have high ability, it is efficient to fill all positions from this group. Similarly, when $k > \tilde{k}$, elite capacity is so large that some of those with the lowest attainable training level (who also have low ability) must be selected and all those with higher levels of training are selected with certainty. In both of these cases, monotonic and group blind policies are optimal. Since optimal policies are monotonic, they will also be equilibrium policies, with no individual actually choosing to underinvest. And allowing for commitment to a policy again changes nothing, the screener will simply commit to the optimal policy.

If $k \in (\bar{k}, \tilde{k})$ and (9) and (10) hold, there may be incentives to underinvest. Some or all of those who can attain t_{lh} might choose the lower training level t_{hl} to increase their likelihood of selection. This in turn, affects candidate quality at t_{hl} . Using η_i to denote the proportion of high ability candidates at t_{hl} for group i , expected performance at this training level is

$$\eta_i \phi(a_h, t_{hl}) + (1 - \eta_i) \phi(a_l, t_{hl}).$$

With no underinvestment, $\eta_i = 1$ and expected performance at t_{hl} is at its highest possible value. At the other extreme, if all individuals who could attain t_{lh} choose t_{hl} , η_i takes its lowest feasible value, which we denote by λ_i . This is the share of high ability types at t_{hl} under complete pooling:

$$\lambda_i = \frac{(1 - q_i)(1 - F(a_l))}{(1 - q_i)(1 - F(a_l)) + q_i F(a_l)}.$$

Note that if $q_1 < q_2$, then we must have $\lambda_1 > \lambda_2$. Recall that when (9) and (10) both hold, we have

$$\phi(a_h, t_{hl}) > \phi(a_l, t_{lh}) > \phi(a_l, t_{hl}). \quad (11)$$

There must therefore exist $\mu \in (0, 1)$ such that:

$$\mu \phi(a_h, t_{hl}) + (1 - \mu) \phi(a_l, t_{hl}) = \phi(a_l, t_{lh}). \quad (12)$$

This is the proportion of high ability individuals in a pool with training t_{hl} that would result in the same expected performance as low ability individuals with the higher training

level t_{lh} . The relative values of λ_i and μ play a crucial role in determining the structure of equilibrium.

An equilibrium is a pair of probability distributions $\pi_i(t)$ chosen by the screener, along with level of $\eta_i \in [\lambda_i, 1]$. When λ_i is larger than μ , equilibrium selection for group i will be non-monotonic, and all candidates whose attainable training is at the two intermediate training levels will pool at t_{hl} . This is formally stated below.

Proposition 3. *Suppose that $k \in (\bar{k}, \tilde{k})$ and (9) and (10) both hold. Then, in equilibrium: (i) $\lambda_i > \mu$ implies $\pi_i(t_{lh}) = 0$, and (ii) $\lambda_i < \mu$ implies $\pi_i(t_{hl}) = \pi_i(t_{lh})$.*

Proposition 3 describes equilibrium allocations for each group, taken separately, but does not tell us whether the resulting allocation is group-blind.

To examine this, let $T = \{t_{hl}, t_{lh}\}$ denote the set of intermediate training levels, and let $\pi_i(T)$ denote the probability that those in this set will be selected conditional on membership of group i . Then we have:

Proposition 4. *Suppose that $k \in (\bar{k}, \tilde{k})$ and (9) and (10) both hold. Then:*

- a) If $\lambda_1 > \mu$, $\pi_1(T) > \pi_2(T)$ in equilibrium. Furthermore, if $\pi_1(T) < 1$ then $\pi_2(T) = 0$.*
- b) If $\lambda_1 < \mu$, $\pi_i(t_{hl}) = \pi_i(t_{lh})$ for each i in equilibrium, and $\pi_1(T) = \pi_2(T)$ is consistent with equilibrium.*

In the first case, the probability of selection is strictly greater for those in the disadvantaged group at intermediate training levels. All low ability high resource types in the disadvantaged group underinvest, pooling with those having high ability and low resource access. In the advantaged group such pooling may be partial if $\lambda_2 < \mu$, in which case low ability high resource types distribute themselves across the two intermediate training levels in such a manner as to equalize performance, with $\eta_2 = \mu$.

In the second case, individuals in *both* groups distribute themselves across the two intermediate training levels in such a manner as to make expected performance equal to $\phi(a_l, t_{lh})$. This allows the equilibrium selection policy to be blind to group identity. Although the policy is (weakly) monotonic, there is underinvestment in equilibrium.

Proposition 4 implies that the equilibrium selection policy will favor the disadvantaged group for some values of elite capacity, provided that $\lambda_1 > \mu$. We illustrate this below, building on Example 2.

Example 3. Suppose that all specifications are as in Example 2. It may be verified that $\lambda_1 = 27/34 > 1/8 = \mu$. In equilibrium, $\pi_1(t_{hh}) = \pi_2(t_{hh}) = 1$, $\pi_1(t_{hl}) = 4/17$, and $\pi_i(t) = 0$ elsewhere. Among those selected in equilibrium, 55% are from the disadvantaged group.

In this example, all those at the highest training level are selected with certainty. These candidates constitute 6% of the total population, with 1.5% coming from the disadvantaged group. The remaining 4% are drawn exclusively from the disadvantaged group, at training level t_{hl} . Those in the disadvantaged group with low ability and high resources underinvest strategically to pool with the high ability low resource types at this training level. In the advantaged group, only those at the highest training level are selected. This results in their underrepresentation among selected candidates in equilibrium.¹⁰

Since underinvestment lowers performance, one may ask whether a screener who can commit to a policy could obtain a higher equilibrium payoff. Indeed, it is easily seen that a policy that treats the two intermediate training levels symmetrically (as if they were pooled) can achieve this. In this case there is no incentive for underinvestment, so overall performance is higher. But there is no change in the probability of selection—all those at the highest training level are selected (regardless of group), and all those in the disadvantaged group at intermediate training levels are selected with the same positive probability identified in the example. Again, the disadvantaged group is overrepresented in equilibrium.

The point here is not that overrepresentation of a disadvantaged group is likely or even possible in most cases, only that such overrepresentation ought not to be taken as *prima facie* evidence that meritocratic goals are not being pursued.

6 Discussion

Any practical implementation of a meritocratic ideal must be based on observable credentials. In popular discourse, merit-based allocations are inconsistent with selection based on markers of group identity, such as ethnicity, gender, or religion, since these are not intrinsically related to performance. It is also believed that selection based on merit ought to satisfy a monotonicity property; those with higher values of performance-related attributes, such as test scores or course grades, should have precedence over those with

¹⁰See Sethi and Somanathan (2023) for the full range of representation outcomes as elite capacity varies in this example.

lower values. In this paper we have challenged the general nature of both of these claims. Our arguments reveal that the usual framing of the problem of affirmative action—as a trade-off between performance and representation—is misleading.

We show that selection rules that maximize organizational performance can be both group-conscious and non-monotonic in observable credentials. Whether or not they are in fact so, depends on the joint distribution of observable and unobservable attributes (which we term training and ability respectively), their relative importance in determining performance, and the fraction of the eligible pool of candidates that is selected. An implication of these results is that without these pieces of information, we cannot reasonably infer the weights that organizations attach to diversity based on the degree of representation among their students or workers.

It is often assumed that in the face of group inequality, market incentives will lead to limited mixing between groups. Our paper shows that this need not be true. In fact, misconceptions related to the nature of meritocratic allocations may result in legal and institutional obstacles to representation that could potentially improve performance.

A Appendix

A.1 Proofs

We begin with a preliminary result that is required for the proof of Proposition 1. Lemma 1 appears as Equation (4) in the text and states that at any training level at which applicants could have both high and low resource access, Group 1 applicants have higher expected performance than those from Group 2.

Lemma 1. $E(p_1|t) > E(p_2|t)$ if $t \leq t^*$.

Proof of Lemma 1. Note that for any $t' \leq t \leq t^*$,

$$\begin{aligned}\Pr[t' \leq t | r = r_l] &= F(\alpha_l(t)) \\ \Pr[t' \leq t | r = r_h] &= F(\alpha_h(t))\end{aligned}$$

These are the distribution functions for training for low and high resource individuals respectively. The corresponding density functions for training are $f(\alpha_l(t))\alpha'_l(t)$ and $f(\alpha_h(t))\alpha'_h(t)$. Therefore, using Bayes' rule, when $t \leq t^*$, the probability that an individual in group i has high resources conditional on training t is given by

$$\gamma_i(t) = \frac{q_i f(\alpha_h(t))\alpha'_h(t)}{q_i f(\alpha_h(t))\alpha'_h(t) + (1 - q_i)f(\alpha_l(t))\alpha'_l(t)}. \quad (13)$$

It is easily verified that $q_1 < q_2$ implies $\gamma_1(t) < \gamma_2(t)$ for any $t \leq t^*$. Using the expression for expected performance in (3), and the fact that $\phi(\alpha_h(t), t) < \phi(\alpha_l(t), t)$, we get

$$E(p_1|t) > E(p_2|t).$$

□

Proposition 1: The performance-maximizing policy is monotonic in training and group-blind if $k < \hat{k}$, and group-sighted and favors the disadvantaged group if $k > \hat{k}$.

Proof of Proposition 1. We start by showing that for elite capacity below $\hat{k}$, the performance-maximizing policy is monotonic in training and group-blind.

Recall that $\hat{t}_i$ is defined as the training level at which the expected performance of a candidate with high resource access is the same as the highest expected performance for

a Group i candidate at any training level below t^* , and $\hat{k}$ is the proportion of the applicant population (aggregated across both groups) that has training level at or above $\hat{t}_1$. By Lemma 1, $\hat{t}_2 < \hat{t}_1$. If $k \leq \hat{k}$ then there exists $t' \geq \hat{t}_1$ such that

$$k = \sum_{i=1}^2 s_i q_i (1 - F(\alpha_h(t'))).$$

That is, elite capacity k can be filled by selecting all candidates in $[t', 1]$, regardless of group membership, and rejecting all other candidates. This is a monotonic and group-blind policy. To see that it is also performance maximizing, note that since $t^* < \hat{t}_2 < \hat{t}_1 \leq t'$, all those with training above t' (regardless of group) have higher productivity than all those with training below t' .

Lemma 1, together with the fact that $p_1 = p_2$ for $t > t^*$, implies that the advantaged group can never be favored under a performance-maximizing policy. To see that the disadvantaged group must be favored if $k > \hat{k}$, note first that in this case the set of training levels $t \in [0, t^*)$ such that $\pi_1(t) = 1$ has positive measure. If there is no t in this interval with $\pi_2(t) > 0$ then clearly the disadvantaged group is favored. If there is a $t < t^*$ for which a candidate from Group 2 is selected, then we know that the lowest level of expected performance among those selected must be equalized across groups:

$$\min_{\pi_1(t)=1} E(p_1|t) = \min_{\pi_2(t)=1} E(p_2|t). \quad (14)$$

The above condition must hold, otherwise expected performance could be raised by replacing some members of one group with an equal measure from the other. But from Lemma 1, this is only possible if there is some set of training levels $t < t^*$ such that $\pi_1(t) > 0 = \pi_2(t)$. That is, the disadvantaged group is favored. $\square$

Proposition 2: If the performance function is given by $\phi(a, t) = \beta a + (1 - \beta)t$, the threshold level of elite capacity $\hat{k}$ below which the performance-maximizing allocation is monotonic and group-blind is decreasing in β .

Proof of Proposition 2. Fix any group i , set $\gamma = \gamma_i(t^*)$, and define $H(\beta, t)$ as follows

$$\begin{aligned} H(\beta, t) &= \gamma \phi(\alpha_h(t^*), t^*) + (1 - \gamma) \phi(\alpha_l(t^*), t^*) - \phi(\alpha_h(t), t) \\ &= \gamma(\beta \alpha_h(t^*) + (1 - \beta)t^*) + (1 - \gamma)(\beta \alpha_l(t^*) + (1 - \beta)t^*) - (\beta \alpha_h(t) + (1 - \beta)t) \\ &= \beta(\gamma \alpha_h(t^*) + (1 - \gamma) \alpha_l(t^*)) + (1 - \beta)t^* - (\beta \alpha_h(t) + (1 - \beta)t). \end{aligned}$$

Note that $\hat{t}_i$ is a solution for t to $H(\beta, t) = 0$ and that $\partial H/\partial t$ is everywhere nonzero. This implicitly defines a function $\hat{t}_i(\beta)$. Using the Implicit Function Theorem, we obtain

$$\frac{d\hat{t}_i}{d\beta} = -\frac{\partial H/\partial \beta}{\partial H/\partial t} = \frac{\gamma\alpha_h(t^*) + (1-\gamma)\alpha_l(t^*) - \alpha_h(\hat{t}_i) + \hat{t}_i - t^*}{\beta\alpha'_h(\hat{t}_i) + (1-\beta)}$$

Since $\hat{t}_i > t^*$, the above expression is also positive provided that

$$\gamma\alpha_h(t^*) + (1-\gamma)\alpha_l(t^*) > \alpha_h(\hat{t}_i). \quad (15)$$

To see that this must be the case, note that by definition,

$$\gamma(\beta\alpha_h(t^*) + (1-\beta)t^*) + (1-\gamma)(\beta\alpha_l(t^*) + (1-\beta)t^*) = \beta\alpha_h(\hat{t}_i) + (1-\beta)\hat{t}_i.$$

Hence

$$\beta(\gamma\alpha_h(t^*) + (1-\gamma)\alpha_l(t^*) - \alpha_h(\hat{t}_i)) = (1-\beta)(\hat{t}_i - t^*) > 0,$$

which implies (15). Hence $\hat{t}_i$ is increasing in β . The claim then follows from the definition of $\hat{k}$. $\square$

Proposition 3: Suppose that $k \in (\bar{k}, \tilde{k})$ and (9) and (10) both hold. Then, in equilibrium: (i) $\lambda_i > \mu$ implies $\pi_i(t_{lh}) = 0$, and (ii) $\lambda_i < \mu$ implies $\pi_i(t_{hl}) = \pi_i(t_{lh})$.

Proof of Proposition 3. First consider the case $\lambda_i > \mu$, which implies

$$\lambda_i\phi(a_h, t_{hl}) + (1-\lambda_i)\phi(a_l, t_{hl}) > \phi(a_l, t_{lh})$$

from (11–12). Hence $\pi_i(t_{lh}) > 0$ implies $\pi_i(t_{hl}) = 1$. But this means that all those who could attain t_{lh} without underinvesting will choose to attain $t_{hl} < t_{lh}$ instead. Thus $\pi_i(t_{lh}) = 0$.

Next consider $\lambda_i < \mu$, in which case

$$\lambda_i\phi(a_h, t_{hl}) + (1-\lambda_i)\phi(a_l, t_{hl}) < \phi(a_l, t_{lh})$$

from (11–12). If $\pi_i(t_{hl}) < \pi_i(t_{lh})$, then there will be no underinvestment, in which case the screener will prefer those at t_{hl} , contradicting $\pi_i(t_{hl}) < \pi_i(t_{lh})$. So we must have $\pi_i(t_{hl}) \geq \pi_i(t_{lh})$. If this inequality is strict, then all those who could attain t_{lh} without underinvestment will choose the lower training level t_{hl} instead. In this case expected performance at t_{hl} will fall below that at t_{lh} , contradicting $\pi_i(t_{hl}) > \pi_i(t_{lh})$. Thus $\pi_i(t_{hl}) = \pi_i(t_{lh})$, as claimed. $\square$

Proposition 4: Suppose that $k \in (\bar{k}, \tilde{k})$ and (9) and (10) both hold. Then:

- a) If $\lambda_1 > \mu$, $\pi_1(T) > \pi_2(T)$ in equilibrium. Furthermore, if $\pi_1(T) < 1$ then $\pi_2(T) = 0$.
- b) If $\lambda_1 < \mu$, $\pi_i(t_{hl}) = \pi_i(t_{lh})$ for each i in equilibrium, and $\pi_1(T) = \pi_2(T)$ is consistent with equilibrium.

Proof of Proposition 4. Suppose $k \in (\bar{k}, \tilde{k})$ and $\lambda_1 > \mu$. We first consider the case $\lambda_2 > \mu$. Suppose that types (a_l, r_h) in both groups underinvest and choose t_{hl} while all other types invest at the highest feasible level. Then a selection policy with $\pi_i(t_{ll}) = \pi_i(t_{lh}) = 0$ and $\pi_i(t_{hh}) = 1$ is a best response if either $\pi_1(t_{hl}) = 1 > \pi_2(t_{hl})$, or $\pi_1(t_{hl}) > 0 = \pi_2(t_{hl})$. To see this, note that if $\pi_1(t_{hl}) < 1$ and $\pi_2(t_{hl}) > 0$ then expected performance can be raised by shifting probability from $\pi_2(t_{hl})$ to $\pi_1(t_{hl})$. Under the proposed selection policy, no individual can profitably deviate by changing her investment choice.

If $k \in (\bar{k}, \tilde{k})$ and $\lambda_1 > \mu > \lambda_2$, there is an equilibrium in which all types (a_l, r_h) in the disadvantaged group underinvest and choose t_{hl} while those of type (a_l, r_h) in the advantaged group distribute themselves across the training levels t_{hl} and t_{lh} in such a manner as to make the expected performance equal at both training levels. Then a selection policy with $\pi_i(t_{ll}) = \pi_1(t_{lh}) = 0$ and $\pi_i(t_{hh}) = 1$ is a best response if either $\pi_1(t_{hl}) = 1 > \pi_2(t_{hl}) = \pi_2(t_{lh})$, or $\pi_1(t_{hl}) > 0 = \pi_2(t_{hl}) = \pi_2(t_{lh})$. To see this, note that if $\pi_1(t_{hl}) < 1$ and either $\pi_2(t_{hl})$ or $\pi_2(t_{lh}) > 0$ then expected performance can be raised by shifting probability from $\pi_1(t_{hl})$ from either $\pi_2(t_{hl})$ or $\pi_2(t_{lh})$. Under the proposed selection policy, no individual can profitably deviate by changing her investment choice.

Finally, suppose that $k \in (\bar{k}, \tilde{k})$ and $\mu > \lambda_1$. Then there is an equilibrium in which all types (a_l, r_h) in both groups distribute themselves across the training levels t_{hl} and t_{lh} in such a manner as to make the expected performance equal to μ at both training levels. Then a selection policy with $\pi_i(t_{ll}) = 0$ and $\pi_i(t_{hh}) = 1$ is a best response if $\pi_i(t_{hl}) = \pi_i(t_{lh})$. Under the proposed selection policy, no individual can profitably deviate by changing her investment choice. $\square$

A.2 Continuous Resource Distributions

While the case of two resource levels makes our reasoning transparent, this is not required to generate either the non-monotonicity of expected performance in training, or selection

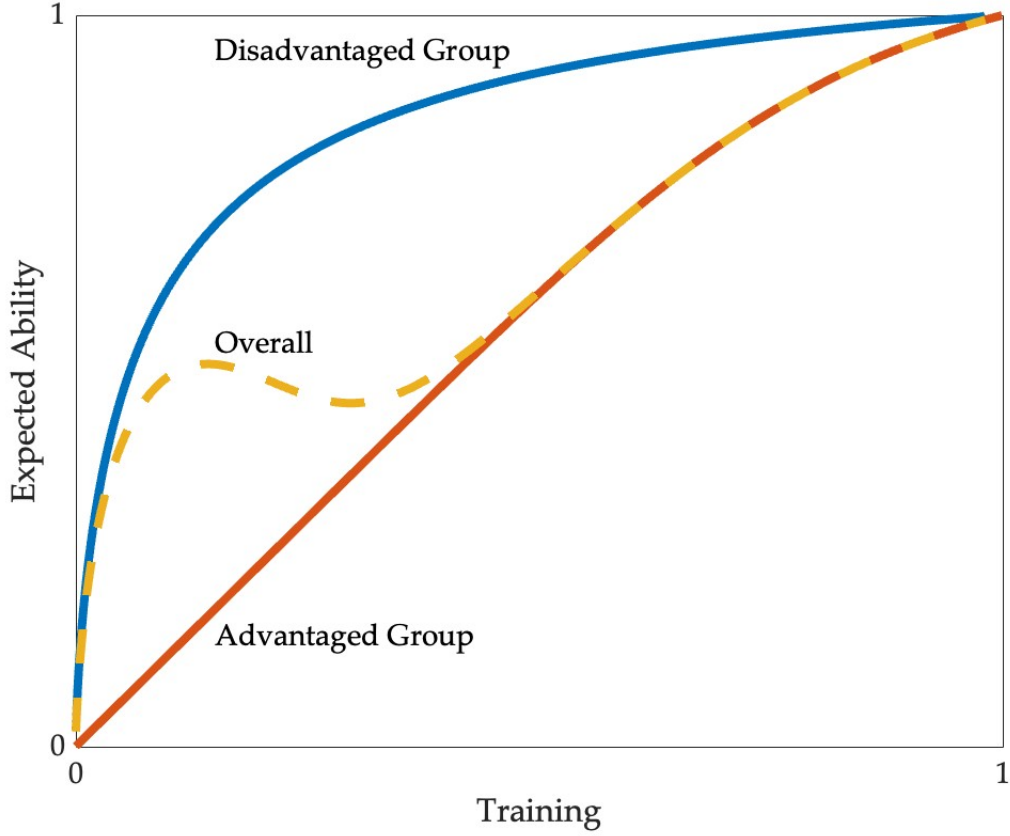


Figure 4: Expected Ability by Group and Overall.

policies that favor the disadvantaged group. This can be seen by considering a case with continuous resource distributions.

Suppose that ability and resources are independently distributed on $[0,1]$, with the ability distribution being uniform. The two resource distributions are both beta distributions with shape parameters (2,10) in the disadvantaged group and (10,2) in the advantaged group. The training function is $\tau(a, r) = ar$ and there is no strategic underinvestment. The two groups are of equal size.

Conditional on training t , the expected ability in each group, as well as that in the population as a whole, is shown in Figure 4. The disadvantaged group has higher expected ability (conditional on training) at all training levels, and hence also has higher expected performance conditional on training, regardless of the performance function (as long as performance depends on both ability and training). So the disadvantaged group will be favored under a performance-maximizing selection policy (regardless of elite capacity) and the imposition of group-blind selection will lower performance.

Although expected ability is increasing in training within any group, this is not the case for the population as a whole. Whether or not expected performance in the population as a whole is likewise non-monotonic will depend on the degree to which ability rather than training matters for performance. If ability has sufficient weight, the overall expected performance will also inherit the non-monotonicity of expected ability. In this case the imposition of a group-blind policy will result in non-monotonic selection to maximize performance.

This example also shows that expected performance can be non-monotonic *within* groups. For instance, the overall population here can itself be considered as a single group, with a bimodal resource distribution obtained by simply aggregating the two component distributions.

References

- Dennis J. Aigner and Glen G. Cain. Statistical theories of discrimination in labor markets. Industrial and Labor Relations Review, pages 175–187, 1977.
- Richard Arneson. Equality of Opportunity. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Summer 2015 edition, 2015.
- Joshua Aronson and Claude M. Steele. Stereotypes and the fragility of academic competence, motivation, and self-concept. Handbook of competence and motivation, pages 436–456, 2005.
- Kenneth J. Arrow. The theory of discrimination. In Orley Ashenfelter and Albert Rees, editors, Discrimination in Labor Markets. Princeton, NJ: Princeton University Press, 1973.
- Matt Cavanagh. Against equality of opportunity. Clarendon Press, 2002.
- Dario Cestau, Dennis Epple, and Holger Sieg. Admitting students to selective education programs: Merit, profiling, and affirmative action. Journal of Political Economy, 125(3):761–797, 2017.
- Jimmy Chan and Erik Eyster. Does banning affirmative action lower college student quality? American Economic Review, 93(3):858–872, 2003.
- Stephen Coate and Glenn C. Loury. Will affirmative-action policies eliminate negative stereotypes? American Economic Review, 83(5):1220–1240, 1993.
- Bradford Cornell and Ivo Welch. Culture, information, and screening discrimination. Journal of Political Economy, pages 542–571, 1996.
- Steven N. Durlauf. Affirmative action, meritocracy, and efficiency. Politics, Philosophy & Economics, 7(2):131–158, 2008.
- Roland G. Fryer and Glenn C. Loury. Valuing diversity. Journal of Political Economy, 121(4):747–774, 2013.
- Roland G. Fryer, Glenn C. Loury, and Tolga Yuret. An economic analysis of color-blind affirmative action. Journal of Law, Economics, and Organization, 24(2):319–355, 2008.

- Daniel Markovits. The Meritocracy Trap: How America's Foundational Myth Feeds Inequality, Dismantles the Middle Class, and Devours the Elite. Penguin Press, 2019.
- Edmund S. Phelps. The statistical theory of racism and sexism. American Economic Review, 62(4):659–661, 1972.
- Debraj Ray and Rajiv Sethi. A remark on color-blind affirmative action. Journal of Public Economic Theory, 12(3):399–406, 2010.
- C.D.C. Reeve. Politics: A New Translation. Hackett Publishing, 2017.
- John E Roemer. Equality of Opportunity. Harvard University Press, 1998.
- Michael J Sandel. The Tyranny of Merit: What's Become of the Common Good? Farrar, Straus and Giroux, 2020.
- TM Scanlon. Why Does Inequality Matter? Oxford University Press, 2018.
- Amartya Sen. Merit and justice. In Kenneth Arrow, Samuel Bowles, and Steven N Durlauf, editors, Meritocracy and Economic Inequality, pages 5–16. Princeton University Press, 2000.
- Rajiv Sethi and Rohini Somanathan. Meritocracy and representation. Journal of Economic Literature, 61(3):941–957, 2023.
- Gregory M Walton, Steven J Spencer, and Sam Erman. Affirmative meritocracy. Social Issues and Policy Review, 7(1):1–35, 2013.