Discrimination as domination: A theory of discriminatory institutions*

James P. Choy

January 8, 2024

Abstract

Institutions in some societies force employers to discriminate. I show that institutionalized discrimination optimally takes the form of *domination*. Discrimination as domination sorts different social groups into different tasks, with an elasticity of substitution between tasks performed by different groups that is as low as possible. The oppressed group receives a lower wage than the dominant group, but the oppressed group is not completely excluded from the labor market. The wage of the dominant group increases in the size of the oppressed group. I apply my theory to explain patterns of discrimination in apartheid South Africa and other discriminatory societies.

Keywords: Discrimination, apartheid, tasks

JEL Classification Numbers: J71, P48

*Corresponding author: James P. Choy, University of York. E-mail: james.choy@york.ac.uk.

1 Introduction

Most economic theories of discrimination describe discrimination practiced by individuals. Individuals may discriminate because of their preferences, as in theories of taste-based discrimination, or because of their beliefs, as in theories of statistical discrimination. However, some of the most important forms of discrimination are imposed not by individuals, but rather collectively by members of a politically powerful social group (the dominant group) against members of a less powerful social group (the oppressed group). I refer to discrimination that is imposed collectively as institutionalized discrimination. Institutionalized discrimination can be enforced by the law and the formal institutions of the state, or by informal institutions and social norms, often backed up by the threat of extra-legal violence. Some of the most notorious examples of societies that have institutionalized discrimination include apartheid South Africa, the US South under Jim Crow, and Nazi Germany.

Many previous authors have noted that discriminatory institutions function by reserving certain jobs for members of the dominant group. In apartheid South Africa, many jobs were reserved by law for Whites. In the Jim Crow South, discrimination was for the most part not enforced by law. However, social norms informally reserved many jobs for Whites, and employers and workers who violated these social norms could face violent consequences, inflicted either by spontaneously formed mobs or by more organized groups such as the Ku Klux Klan. Job reservations benefit workers from the dominant group by increasing their wages. But which jobs do discriminatory institutions reserve? In other words, what pattern of discrimination is optimal for workers in the dominant group? And can a theory of optimal discriminatory institutions help to explain observed patterns of discrimination? These questions, which have not been addressed previously, are the questions that motivate this paper.

I construct a model in which there are two social groups, a dominant group and an oppressed group. Workers from each group can choose to work in any of a number of different tasks. The dominant group may use its political power to impose labor market regulations reserving some subset of the available tasks for members of the dominant group. Task reservation increases the wage for reserved tasks above the free market wage, and reduces the wage for unreserved tasks below the free market wage, even though all tasks require the same level of skill. Thus, task reservation breaks the law of one price for tasks proposed by Acemoglu and Autor (2011), which states that tasks that require the same level of skill receive the same wage.

By choosing the set of reserved tasks appropriately, the dominant group can choose both the size of the set of reserved tasks and the elasticity of substitution between reserved and unreserved tasks. I show that the dominant group optimally chooses the elasticity of substitution between reserved and unreserved tasks to be as low as possible. It is important for this result that the dominant group can choose both the size of the set of reserved tasks and the elasticity of substitution between reserved and unreserved tasks. If the size of the set of reserved tasks is exogenously fixed, then in general it is not optimal for the dominant group to set the elasticity of substitution between reserved and unreserved tasks as low as possible. Under optimal discrimination, the ratio of reserved to unreserved tasks is greater than the ratio of dominant group to oppressed group workers. However, if the minimum feasible elasticity of substitution between reserved and unreserved tasks is not too large, then not all tasks are reserved.

Next I show how wages under discrimination vary depending on the relative sizes of the two groups. I show that under optimal discrimination, the wage of workers in the dominant group is increasing in the size of the oppressed group. This result also depends on the assumption that the set of reserved tasks is chosen optimally. If the set of reserved tasks is exogenously fixed, and if the fixed elasticity of substitution between reserved and unreserved tasks is sufficiently large, then the dominant group wage is strictly decreasing in the size of the oppressed group.

Together, these results describe a form of discrimination that I call discrimination as domination. Under discrimination as domination, members of different groups are forcibly sorted into economic roles that are as different as possible, where "different" is defined in the economic sense of there being a low elasticity of substitution between different roles. These roles are hierarchically ranked in the sense that members of the oppressed group receive lower wages than members of the dominant group. Although members of the oppressed group are forced into an economic role that is constrained and inferior relative to the role played by members of the dominant group, members of the oppressed group are not completely excluded from the labor market. Members of the dominant group exploit members of the oppressed group, and so members of the dominant group benefit from increasing the size of the oppressed group.

In the literature, the paper most closely related to mine is Bergmann (1971). Bergmann presents a model which is formally equivalent to mine, except that in Bergmann's model the set of reserved tasks is exogenously fixed. Since the set of reserved tasks is fixed in Bergmann's model, Bergmann does not discuss what set of reserved tasks would be optimal for the dominant group. The statistical discrimination models of Norman (2003) and Moro and Norman (2004) can also be interpreted as models in which there is an exogenously fixed set of tasks that can be subject to discrimination.

The main alternative theory of institutionalized discrimination is that institutionalized discrimination is designed to increase the capital-labor ratio or land-labor ratio for workers in the dominant group, by excluding workers from the oppressed group from access to the relevant factor of production. I refer to this theory as the theory of factor-based institutionalized discrimination. Factor-based institutionalized discrimination is modeled by Krueger (1963), Porter (1978), and Lundahl (1982). The theory of factor-based institutionalized discrimination has also been developed informally in works such as Hutt (1964) and Lipton (1985).[1]

Like discrimination as domination, factor-based institutionalized discrimination operates by reserving a subset of available tasks for members of the dominant group. However, the set of reserved tasks under factor-based institutionalized discrimination is different from the set of reserved tasks under discrimination as domination. Specifically, under factor-based institutionalized discrimination, it is optimal to exclude the oppressed group from the labor market completely, since excluding the oppressed group completely maximizes the amount of the non-labor factor of production that is available to the dominant group. I show that discrimination as domination is optimal if the elasticity of substitution between labor and the non-labor factor of production is sufficiently large, while factor-based institutionalized discrimination may be optimal if the elasticity of substitution between labor and the non-labor factor of production is small. I argue that this result helps to explain why discrimination as domination seems to appear primarily in relatively modern, industrialized societies, while factor-based institutionalized discrimination also appears in historical and less-developed societies.

One of the most effective ways of excluding members of one social group from the labor market completely is by physically removing members of that group from society. Thus, the theory of factor-based institutionalized discrimination can be interpreted as a theory of ethnic cleansing or genocide.[2] In contrast, under discrimination as domination, the dominant group benefits from increasing the size of the oppressed group. I argue that the distinction between factor-based institutionalized discrimination and discrimination as domination helps to explain disagreements between factions in the apartheid South African government about whether to increase or decrease the size of the Black population in White areas of South Africa.

I also discuss four other distinctive predictions of my model. First, my model predicts that discrimination can increase the prevalence of socially heterogeneous firms relative to the free market. This result contrasts with the argument of Becker (1957) that discrimination reduces the prevalence of socially heterogeneous firms relative to the free market. Second, my model suggests that the effect of discrimination on the return to capital may be relatively small. This result contrasts with the argument of Lipton (1985) that institutionalized discrimination significantly reduces the return to capital and that capital owners therefore form a natural anti-discrimination political constituency. Third, my model predicts that the dominant group may be willing to expend resources to increase the effective labor supply of the oppressed group, for example

[1] Fang and Norman (2006) also develop a theory of some of the consequences of institutionalized discrimination. However, they simply assume that institutionalized discrimination exists and takes a particular form, they and do not explain why institutionalized discrimination exists or what purpose it serves. Small and Pager (2020) argue that existing economic models do not describe institutionalized discrimination, although their concept of institutionalized discrimination is different from mine and does not include the idea of discrimination as domination. The field of stratification economics (Darrity, 2005, 2022; Chelwa et al., 2022) studies a concept of discrimination that is close to mine in some ways. However, this field is primarily empirical and does not include any formal models of discrimination as domination as far as I am aware.

[2] Esteban et al. (2015) develop a theory of genocide along these lines.

by investing in the human capital of the oppressed group. This result helps to explain the fact that the South African apartheid state invested a non-zero quantity in Black education and health care, a fact that Seekings and Nattrass (2005) present as a puzzle. Finally, my model predicts that, since the set of reserved tasks is chosen optimally in response to economic conditions, the set of reserved tasks may change quickly in response to changing economic conditions. This result contrasts with the argument that the set of reserved tasks is determined by tastes for discrimination among members of the dominant group, presented for example by Hurst et al. (2022). If the set of reserved tasks is determined by tastes for discrimination, then the set of reserved tasks is likely to change slowly if at all in response to changing economic conditions, since tastes are likely to change slowly if at all in response to economic conditions. I discuss all of these predictions in the context of apartheid South Africa.

I extend my model by discussing how discrimination as domination is enforced. I argue that the enforcement of discrimination can be very costly, and in particular that enforcing discrimination is generally more costly than collecting taxes. Nevertheless, the dominant group may prefer redistribution through discrimination to redistribution through taxation. Since the costs of enforcing discrimination are unproductive, discriminatory institutions are inefficient relative to non-discriminatory institutions. I argue that this form of inefficiency is novel and differs in particular from the forms of inefficiency under extractive institutions in the taxonomy presented by Acemoglu (2006). I present evidence suggesting that this form of inefficiency was a first-order contributor to the overall inefficiency of the South African economy under apartheid.

Finally, I present examples of discrimination as domination in societies other than South Africa. I discuss the US South under Jim Crow, contemporary Saudi Arabia, and policies relating to illegal immigrants in many developed societies. I argue that all of these sets of institutions have features broadly consistent with my model of discrimination as domination.

1.1 Relationship to the theory of directed technological change

In my model, labor supplied by the dominant social group and labor supplied by the oppressed social group are in effect different factors of production. The elasticity of substitution between different factors of production plays an important role in the theory of directed technological change. Kamien and Schwartz (1968) show that for fixed factor input levels, total output is increasing in the elasticity of substitution between factors. They argue that a profit maximizing firm may want to invest to increase the elasticity of substitution between different factors, and that this incentive helps to determine the overall direction of technological change. Klump and de La Grandville (2000) use related ideas in a Solow growth model to argue that the rate of economic growth and the steady state level of per capita income are both increasing in the elasticity of substitution between capital and labor. They also conjecture that the elasticity of substitution may increase over time as a consequence of technological change.

In contrast to these arguments, in my model the dominant group has an incentive to invest to reduce the minimum feasible elasticity of substitution between reserved and unreserved tasks, reversing the claims about the direction of technological change suggested by Kamien and Schwartz (1968) and Klump and de La Grandville (2000). The proofs of my results make use of the concept of a normalized constant elasticity of substitution production function, introduced by de La Grandville (1989) and developed by Klump and de La Grandville (2000). This concept has played an important role in recent research on directed technological change, as surveyed for example by Klump et al. (2012). Given the close analogy between my results and results in the theory of directed technological change, the proofs of my results may be of independent interest.

2 Theory

2.1 The production technology

Consider a society which contains two social groups, which I will label "dominant" and "oppressed". The dominant group monopolizes political power, excluding the oppressed group. Each group contains a continuum of workers. Let the measures of the sets of workers in the dominant and oppressed groups be α_d and α_o, respectively. Each worker inelastically supplies one unit of labor to one of a number of tasks. The way labor supplied to particular tasks is combined into aggregate production depends on both the underlying technology and on social institutions. In order to motivate the aggregate production function introduced below, I begin with an example of a specific production technology.

Suppose that the economy consists of a representative firm that is formed from a number of different divisions, all of which work together to produce the final good. The output of each division in turn depends on a variety of tasks performed within each division. Suppose that there are a continuum of divisions in the firm and a continuum of tasks within each division, and that all of these continua have measure 1. Let $\ell(i, j)$ be the quantity of labor supplied to task i within division j. The output $q(j)$ of division j is produced according to the constant elasticity of substitution (CES) function:

$$q(j) = \left[\int_0^1 \ell(i, j)^{(\tau_1 - 1)/\tau_1} di \right]^{\tau_1/(\tau_1 - 1)} \tag{1}$$

Here τ_1 is the elasticity of substitution between tasks within each division. For simplicity I assume that this elasticity is the same across all divisions j.

Aggregate labor supply L is a function of the output of the different divisions and also takes a CES form:

$$L = \left[\int_0^1 q(j)^{(\tau_2 - 1)/\tau_2} dj \right]^{\tau_2/(\tau_2 - 1)} \tag{2}$$

Here τ_2 is the elasticity of substitution between divisions.

The final good is produced using aggregate labor supply L and some other factor of production Z, which could represent physical capital, human capital, or land. The final production function is

$$Y = F(Z, L) \tag{3}$$

The dominant social group can use its political power to determine the form of social institutions. There are two possible social institutions. The first is the free market, under which all workers can choose freely what task to perform. The second is institutionalized discrimination. Under discrimination, some subset of the available tasks is reserved for workers in the dominant group.

All tasks require the same level of skill. Therefore, any worker can perform any task, and so in the free market the wages for all tasks must be equal. This is the law of one price for tasks proposed by Acemoglu and Autor (2011). The functional form of the production function implies that wages for all tasks are equal when the amount of labor applied to every task is the same. Therefore, in the free market the amount of labor applied to every task is $\alpha_d + \alpha_o$. Aggregate production in the free market is then $Y = F(Z, \alpha_d + \alpha_o)$. This production function implies that in the free market workers from different groups are perfect substitutes, regardless of the elasticities of substitution τ_1 and τ_2, and so these elasticities are irrelevant. The division of labor across social groups is indeterminate in the free market equilibrium, as any allocation of workers from different social groups to tasks is consistent with equilibrium as long as the total amount of labor allocated to each task is the same.

Instead of allowing a free market, the dominant group can reserve some subset of tasks for dominant group workers. Suppose that the dominant group wants to reserve a set of tasks with measure R. Consider two ways to do this. First, the dominant group can reserve a measure R of the tasks within each division. If $R \leq \alpha_d/(\alpha_d + \alpha_o)$, then the restriction that oppressed workers cannot perform reserved tasks does not bind, and aggregate production is the same as in the free market. On the other hand, if $R > \alpha_d/(\alpha_d + \alpha_o)$, then the restriction does bind. In this case the wage for reserved tasks is higher than the wage for unreserved tasks, and so all dominant group workers choose reserved tasks, while oppressed group workers can only choose unreserved tasks. Within the sets of reserved and unreserved tasks, the law of one price for tasks still implies that the wages for all tasks are equal and hence that the number of workers assigned to each task within a given set is the same. The production function for each division j then becomes:

$$q(j) = \left[R \left(\frac{\alpha_d}{R} \right)^{(\tau_1 - 1)/\tau_1} + (1 - R) \left(\frac{\alpha_o}{1 - R} \right)^{(\tau_1 - 1)/\tau_1} \right]^{\tau_1/(\tau_1 - 1)} \tag{4}$$

Here α_d/R is the number of dominant group workers per task in the set of reserved tasks, and $\alpha_o/(1 - R)$ is the number of oppressed group workers per task in the set of unreserved tasks.

Since the same measure R of tasks are reserved in each division, production $q(j)$ of each division is the same for all divisions j. Thus aggregate labor supply is:

$$L = \left[R \left(\frac{\alpha_d}{R} \right)^{(\tau_1 - 1)/\tau_1} + (1 - R) \left(\frac{\alpha_o}{1 - R} \right)^{(\tau_1 - 1)/\tau_1} \right]^{\tau_1/(\tau_1 - 1)} \tag{5}$$

Given this form of discrimination, the elasticity of substitution between dominant group workers and oppressed group workers in the aggregate production function is $\sigma = \tau_1$.

Now consider a different way of reserving a measure R of the available tasks. Suppose that instead of reserving a measure R of the tasks within each division, the dominant group chooses a measure R of divisions, and reserves all tasks within these divisions, while leaving all tasks in the other divisions unreserved. In this case, the output of the reserved divisions is:

$$q_r = \frac{\alpha_d}{R} \tag{6}$$

The output of the unreserved divisions is:

$$q_u = \frac{\alpha_o}{1 - R} \tag{7}$$

Aggregate labor supply is:

$$L = \left[R \left(\frac{\alpha_d}{R} \right)^{(\tau_2 - 1)/\tau_2} + (1 - R) \left(\frac{\alpha_o}{1 - R} \right)^{(\tau_2 - 1)/\tau_2} \right]^{\tau_2/(\tau_2 - 1)} \tag{8}$$

Given this form of discrimination, the elasticity of substitution between dominant group workers and oppressed group workers in the aggregate production function is $\sigma = \tau_2$.

The point of this example is that given the underlying production technology, by choosing the set of reserved tasks appropriately the dominant group can choose the size of the set of reserved tasks R and can also decide whether the elasticity of substitution between dominant and oppressed group workers in the aggregate production function is $\sigma = \tau_1$ or $\sigma = \tau_2$.

More generally, it may be possible for the elasticity of substitution between reserved and unreserved tasks to take on many different values depending on the set of reserved tasks. For example, suppose that each task is composed of many different subtasks, with elasticity of substitution τ_3 between subtasks. Then by reserving a measure R of the subtasks that compose each task, the dominant group could set the elasticity of substitution between reserved and unreserved subtasks equal to τ_3. Further refinements of the production technology would yield even more possibilities.

2.2 General setup

As in the previous subsection, I assume that final output is a function of aggregate labor supply L and some other factor of production Z:

$$Y = F(Z, L) \tag{9}$$

Following the example from the previous subsection, I assume that aggregate labor supply L is a CES function of the sizes of the dominant and oppressed groups:

$$L(\alpha_d, \alpha_o, R, \sigma) = \left[R \left(\frac{\alpha_d}{R} \right)^{(\sigma-1)/\sigma} + (1-R) \left(\frac{\alpha_o}{1-R} \right)^{(\sigma-1)/\sigma} \right]^{\sigma/(\sigma-1)} \tag{10}$$

In the example in the previous subsection, by choosing set of reserved tasks, the dominant group could choose R and σ in the aggregate labor supply function L, with a discrete set of possible values of σ. For the remainder of the paper, I abstract from the specific technology suggested in the previous subsection by supposing that the set of possible values of σ is continuous. As will be seen below, the dominant group optimally chooses σ to be as low as possible, so the assumption that σ can vary continuously rather than discretely is mostly innocuous. It may be the case, however, that it is not technologically feasible to choose an elasticity of substitution below some minimum value. Thus the dominant group's choice is subject to the constraint:

$$\sigma \geq \underline{\sigma} \tag{11}$$

Here $\underline{\sigma} \geq 0$ is the minimum feasible value of σ.

Instead of imposing discrimination, the dominant group could allow a free labor market. Following the example in the previous subsection, I assume that aggregate production in the free market is $Y = F(Z, \alpha_d + \alpha_o)$. In the free market, the allocation of workers to tasks is indeterminate as long at the total number of workers in each task is the same.

I assume that $\partial F/\partial Z > 0$, $\partial F/\partial L > 0$, $\partial^2 F/\partial Z^2 < 0$, $\partial^2 F/\partial L^2 < 0$ and $\partial^2 F/\partial Z \partial L > 0$. More substantively, I make the following assumption:

Assumption 1.

$$\frac{\partial}{\partial L}\left(\frac{\partial F}{\partial L}L\right) = \frac{\partial^2 F}{\partial L^2}L + \frac{\partial F}{\partial L} > 0 \tag{12}$$

Assumption 1 states that the total payment to labor, $(\partial F/\partial L)L$, is increasing in L. This assumption is satisfied if the elasticity of substitution between the non-labor factor Z and aggregate labor supply is sufficiently large. For example, if the elasticity of substitution between the non-labor factor and labor is greater than 1, then the labor share of output is increasing in L. By assumption total output also increases in L, so the total payment to labor must be increasing in L. Even if the elasticity of substitution between Z and L is below 1, assumption 1 may be satisfied if the elasticity of total output with respect to aggregate labor supply is sufficiently large. The purpose of assumption 1 will become clear below.

The wage of workers in the dominant group is $w_d = \partial F/\partial \alpha_d = (\partial F/\partial L)(\partial L/\partial \alpha_d)$, and the wage of workers in the oppressed group is $w_o = \partial F/\partial \alpha_o = (\partial F/\partial L)(\partial L/\partial \alpha_o)$. The dominant group chooses R and σ to maximize w_d subject to the constraint that $\sigma \geq \underline{\sigma}$. The assumption that the dominant group maximizes the wage of dominant group workers can be justified by supposing that the majority of members of the dominant group are workers (rather than owners of the non-labor factor Z) and hence that a worker is the median voter within the dominant group.

The following expression for the dominant group wage is useful:

$$w_d(\alpha_d, \alpha_o, R, \sigma) = \frac{\partial F}{\partial L}\left(L\frac{R}{\alpha_d}\right)^{1/\sigma} \tag{13}$$

2.3 Results

2.3.1 The optimal set of reserved tasks

My goal is to characterize the optimal size of the set of reserved tasks R and the optimal elasticity of substitution σ. In order to do this, I will begin by introducing the concept of a normalized CES aggregate labor supply function, first proposed by de La Grandville (1989). The marginal rate of technical substitution (MRTS) between α_d and α_o is

$$MRTS = \frac{\partial L/\partial \alpha_d}{\partial L/\partial \alpha_o} = \left[\frac{R}{(1-R)}\frac{\alpha_o}{\alpha_d}\right]^{1/\sigma} \tag{14}$$

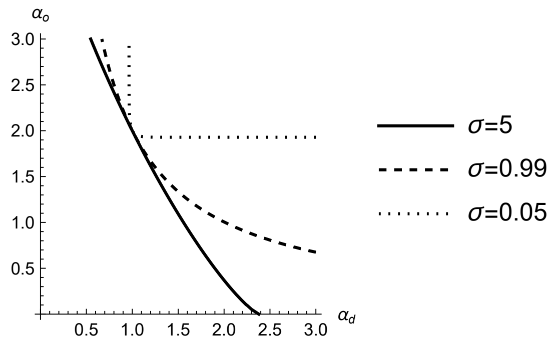
Notice that the MRTS is also equal to the wage ratio w_d/w_o. Let $\bar{\alpha}_d$ and $\bar{\alpha}_o$ be particular values of α_d and α_o and let $\bar{\mu}$ a particular MRTS. Then I can define a family of normalized CES aggregate labor supply that all have MRTS $\bar{\mu}$ at the point $(\bar{\alpha}_d, \bar{\alpha}_o)$, but with different elasticities of substitution σ. More specifically, for each σ, define $R(\bar{\alpha}_d, \bar{\alpha}_o, \bar{\mu}, \sigma)$ to be the value of R such that

$$\left. \frac{\partial L/\partial \alpha_d}{\partial L/\partial \alpha_o} \right|_{\bar{\alpha}_d, \bar{\alpha}_o, R(\bar{\alpha}_d, \bar{\alpha}_o, \bar{\mu}, \sigma)} = \bar{\mu} \tag{15}$$

Define the family of normalized aggregate labor supply functions $\hat{L}$ by

$$\hat{L}(\alpha_d, \alpha_o; R(\bar{\alpha}_d, \bar{\alpha}_o, \bar{\mu}, \sigma), \sigma) = \left[(R(\bar{\alpha}_d, \bar{\alpha_o}, \bar{\mu}, \sigma))^{1/\sigma} \alpha_d^{(\sigma-1)/\sigma} + (1 - R(\bar{\alpha}_d, \bar{\alpha}_o, \bar{\mu}, \sigma))^{1/\sigma} \alpha_o^{(\sigma-1)/\sigma} \right]^{\sigma/(\sigma-1)} \tag{16}$$

Figure 1: A family of normalized CES aggregate labor supply functions



This figure shows the isoquants of three members of the family of CES aggregate labor supply functions normalized to have MRTS $\bar{\mu} = 2$ at the point $(\bar{\alpha}_d, \bar{\alpha}_o)$=(1,2).

Figure 1 shows the isoquants of a family of CES aggregate labor supply functions normalized to have MRTS $\bar{\mu} = 2$ at the point $(\bar{\alpha}_d, \bar{\alpha}_o) = (1, 2)$.

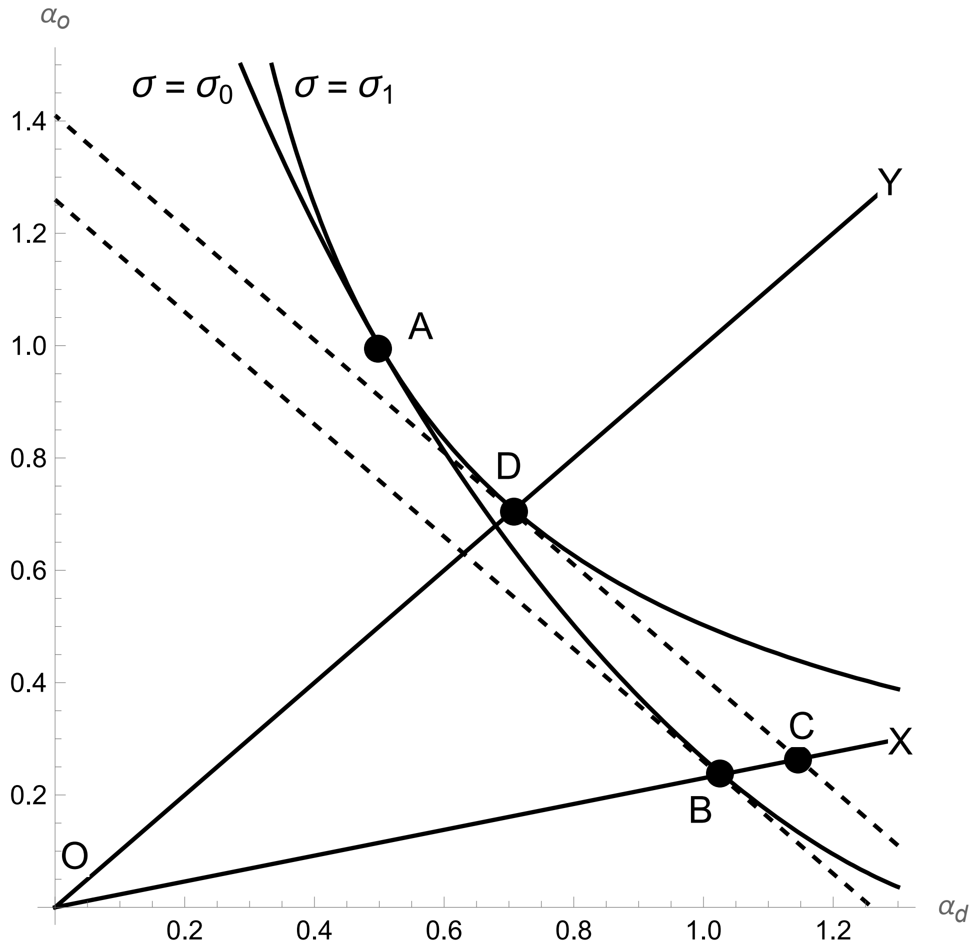
I use these definitions to prove the following lemma:

Lemma 1. *For any $\bar{\alpha}_d$, $\bar{\alpha}_o$, and $\bar{\mu}$ such that $\bar{\mu} > 1$, $\hat{L}(\bar{\alpha}_d, \bar{\alpha}_o; R(\bar{\alpha}_d, \bar{\alpha}_o, \bar{\mu}, \sigma), \sigma)$ is strictly decreasing in σ.*

Proof. The proof of lemma 1 makes use of figure 2, which depicts (α_d, α_o) space. Suppose that the measures of dominant and oppressed group workers are α_d^A and α_o^A. This point is depicted as point A in figure 2. Fix a value of the MRTS $\bar{\mu}$, with $\bar{\mu} > 1$. The figure shows the isoquants of two members of the family of CES aggregate labor supply functions that have slope $\bar{\mu}$ at point A, with elasticities of substitution σ_0 and σ_1, and $\sigma_0 > \sigma_1$.

The ray OX is the set of points where $\alpha_d/(\alpha_d + \alpha_o) = R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0)$. This is the set of points where $\alpha_d/R = \alpha_o/(1 - R)$ for $R = R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0)$. Plugging these equations into (14) shows that the MRTS of the function $\hat{L}(\alpha_d, \alpha_o; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0)$ is equal to 1 at any point (α_d, α_o) on the ray OX. Similarly, the ray OY is the set of points where $\alpha_d/(\alpha_d + \alpha_o) = R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1)$. The MRTS of the aggregate labor supply function $\hat{L}(\alpha_d, \alpha_o; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma_1)$ is equal to 1 at any point (α_d, α_o) on the ray OY. Since the MRTS of both aggregate labor supply functions is greater than 1 at the point A, both the rays OX and OY

Figure 2: Proof of Lemma 1



must be located below A, as depicted in figure 2. In addition, examination of (14) shows that the MRTS is decreasing in σ and increasing in R when the MRTS is greater than 1. Thus, in order to hold the MRTS constant when σ increases, R must also increase. Thus, $R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0) > R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1)$, so the ray OX is located below the ray OY, as depicted in figure 2.

Define (α_d^B, α_o^B) to be the point on ray OX such that:

$$\hat{L}(\alpha_d^A, \alpha_o^A; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) = \hat{L}(\alpha_d^B, \alpha_o^B; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) \tag{17}$$

This point is depicted as point B in figure 2.

Similarly, define (α_d^D, α_o^D) to be the point on ray OY such that

$$\hat{L}(\alpha_d^A, \alpha_o^A; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma_1) = \hat{L}(\alpha_d^D, \alpha_o^D; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma_1) \tag{18}$$

This point is depicted as point D in figure 2.

Finally, define (α_d^C, α_o^C) to be the point where the ray OX intersects the line with slope -1 that goes through point D. This point is depicted as point C in figure 2.

Since $\hat{L}$ is homogeneous of degree 1 in (α_d, α_o), moving outwards along a ray while holding the aggregate labor supply function fixed strictly increases total labor supply. Thus,

$$\hat{L}(\alpha_d^B, \alpha_o^B; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) < \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) \tag{19}$$

On the ray OX, $L = \alpha_d + \alpha_o$ for any σ when $R = R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0)$. Thus,

$$\hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) = \lim_{\sigma \to \infty} \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma) \tag{20}$$

In the limit as σ approaches ∞, L approaches $\alpha_o + \alpha_d$ for any fixed R, and so changing R does not affect total output at the limit. Thus,

$$\lim_{\sigma \to \infty} \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma) = \lim_{\sigma \to \infty} \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma) \tag{21}$$

The line with slope -1 running through point C in figure 2 is an isoquant of the aggregate labor supply function $\lim_{\sigma \to \infty} \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma)$. Since point D is also on this isoquant,

$$\lim_{\sigma \to \infty} \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma) = \lim_{\sigma \to \infty} \hat{L}(\alpha_d^D, \alpha_o^D; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma) \tag{22}$$

On the ray OY, $L = \alpha_d + \alpha_o$ for any σ if $R = R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1)$. Thus,

$$\lim_{\sigma \to \infty} \hat{L}(\alpha_d^D, \alpha_o^D; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma) = \hat{L}(\alpha_d^D, \alpha_o^D; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma_1) \tag{23}$$

Putting (17), (19), (20), (21), (22), (23), and (18) together in order yields:

$$\hat{L}(\alpha_d^A, \alpha_o^A; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) = \hat{L}(\alpha_d^B, \alpha_o^B; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) \tag{24}$$

$$< \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) \tag{25}$$

$$= \lim_{\sigma \to \infty} \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma) \tag{26}$$

$$= \lim_{\sigma \to \infty} \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma) \tag{27}$$

$$= \lim_{\sigma \to \infty} \hat{L}(\alpha_d^D, \alpha_o^D; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma) \tag{28}$$

$$= \hat{L}(\alpha_d^D, \alpha_o^D; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma_1) \tag{29}$$

$$= \hat{L}(\alpha_d^A, \alpha_o^A; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma_1) \tag{30}$$

This completes the proof of lemma 1. □

Lemma 1 shows that when R is varied to hold the wage ratio w_d/w_o fixed, aggregate labor supply L is decreasing in the elasticity of substitution between reserved and unreserved tasks σ. If R is exogenously fixed, this result does not hold. In fact, Kamien and Schwartz (1968) show that for a general CES production function $Y = (aX^{(\sigma-1)/\sigma} + bZ^{(\sigma-1)/\sigma})^{\sigma/(\sigma-1)}$ with factor quantities X and Z and fixed coefficients a and b, total output Y is increasing in σ. This result follows directly from a result in mathematics stating that the generalized mean function $M(t) = (\sum_{i=1}^{n} a_i z_i^t)^{1/t}$ is increasing in t when the coefficients a_i are fixed. (Beckenbach and Bellman, 1961).

Using lemma 1, I can prove the following proposition:

Proposition 1. *Suppose that assumption 1 holds. Then:*

1. *It is optimal to set the elasticity of substitution between reserved and unreserved tasks as low as possible, that is,* $\sigma = \underline{\sigma}$

2. *If* $\underline{\sigma} > 0$ *then the optimal size of the set of reserved tasks* R *satisfies* $R > \frac{\alpha_d}{\alpha_d + \alpha_o}$. *There exists some value* $\bar{\sigma} > 1$ *such that if* $\underline{\sigma} < \bar{\sigma}$, *then* $R < 1$.

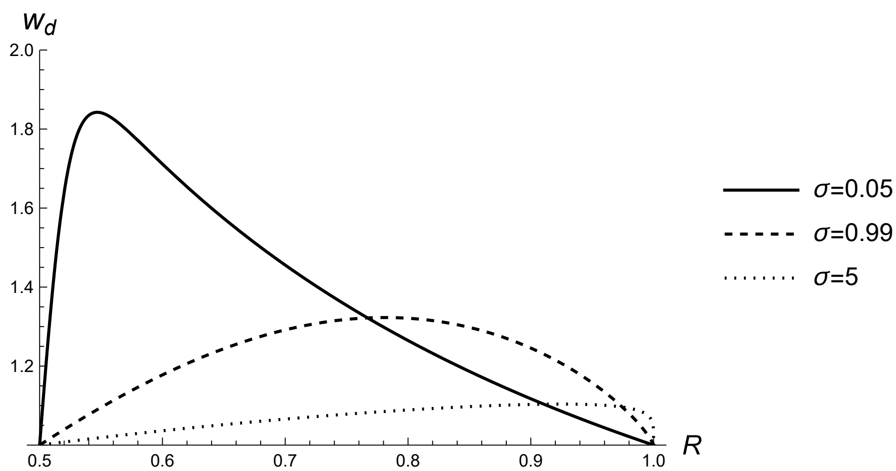
Proof. From lemma 1, decreasing the elasticity of substitution while varying R to hold the wage ratio fixed increases aggregate labor supply L. Assumption 1 states that increasing aggregate labor supply L increases the total payment to labor. If the total payment to labor increases while the wage ratio remains fixed, the wage of the dominant group must increase. Therefore, it is always possible to increase the dominant group wage by reducing the elasticity of substitution, so it is optimal to set the elasticity of substitution as low as possible.

The proof of the second part of proposition 1 is in the appendix. □

The first part of proposition 1 formally expresses the idea that institutionalized discrimination assigns workers from different groups to tasks that are as different as possible, where "different" is defined in the economic sense that the elasticity of substitution between tasks performed by workers from different groups is as low as possible. The second part of proposition 1 states that for all $\underline{\sigma} > 0$ the proportion of reserved tasks is strictly larger than the proportion of dominant group workers within the overall population, so that oppressed group workers are constrained relative to the free market. However, for $\underline{\sigma}$ sufficiently small, the size of the set of reserved tasks is strictly less than 1, so that not all tasks are reserved. Thus, under discrimination as domination, the economic role of the oppressed group is constrained relative to the role of the dominant group. However, the oppressed group still has a role to play, even if this role is constrained, and the oppressed group is not completely excluded from the labor market.

It is important for proposition 1 that both the size of the set of reserved tasks R and the elasticity of substitution between reserved and unreserved tasks σ can vary. If R is fixed, then in general it is not optimal for the dominant group to set σ as low as possible. Figure 3 presents a numerical example illustrating this fact. Let $F(Z, L) = L$, so that the non-labor factor is irrelevant, and let $\alpha_d = \alpha_o = 0.5$. Figure 3 shows the dominant group wage for values of R between 0.5 and 1 and for $\sigma = 0.05$, $\sigma = 0.99$, and $\sigma = 5$. The figure shows that for R greater than approximately 0.77, the dominant group wage is higher for $\sigma = 0.99$ than for $\sigma = 0.05$. Thus when R is fixed at any value greater than 0.77, it is not optimal to set $\sigma = 0.05$ even if it is feasible to do so.

Figure 3: Dominant group wages under discrimination for different values of R and σ



This figure shows the wages of workers in the dominant group w_d when $F(K, L) = L$, $\alpha_d = \alpha_o = 0.5$, for values of R between 0.5 and 1 and for $\sigma = 0.05$, $\sigma = 0.99$, and $\sigma = 5$

2.3.2 Comparing discrimination as domination to factor-based institutionalized discrimination

Suppose now that assumption 1 does not hold, and consider the following alternative assumption:

Assumption 2. *For all* $L \in [\alpha_d, \alpha_d + \alpha_o]$,

$$\frac{\partial}{\partial L}\left(\frac{\partial F}{\partial L} L\right) = \frac{\partial^2 F}{\partial L^2} L + \frac{\partial F}{\partial L} < 0 \tag{31}$$

Assumption 2 states that for all feasible values of L, the total payment to labor is decreasing in aggregate labor supply L. Assumption 2 may hold if the elasticity of substitution between aggregate labor supply L and the non-labor factor of production Z is sufficiently small.

Using assumption 2, I can show the following:

Proposition 2. *Suppose that assumption 2 holds. Then it is optimal for the dominant group to set* $R = 1$. *Any finite value of* σ *is optimal.*

Proof. From (14), the wage ratio w_d/w_o is increasing in R. In the appendix, I show that L is decreasing in R. Therefore, if assumption 2 holds, then increasing R both increases the wage ratio w_d/w_o and increases the total payment to labor, so increasing R must increase the wage w_d. So it is optimal to set R as large as possible, that is, $R = 1$. If $R = 1$ then the wage w_d is the same for all finite values of σ, so any finite value of σ is optimal. □

By setting $R = 1$ and σ equal to some finite value, the dominant group effectively excludes the oppressed group from the labor market completely. This benefits the dominant group by reducing competition from the oppressed group for access to the non-labor factor of production Z. Thus, proposition 2 describes a form of factor-based institutionalized discrimination. One of the most effective ways to exclude an oppressed group from the labor market completely is through ethnic cleansing or genocide, and so proposition 2 also suggests a theory of these phenomena.

Comparing propositions 1 and 2 shows that discrimination as domination is optimal when the elasticity of substitution between labor and the non-labor factor of production is large (and hence assumption 1 holds), while factor-based institutionalized discrimination may be optimal when the elasticity of substitution between labor and the non-labor factor of production is small (and hence assumption 2 is more likely to hold). This result may help to explain why discrimination as domination seems to appear mainly in relatively modern, industrialized societies such as apartheid South Africa and the US South under Jim Crow, while factor-based institutionalized discrimination and related phenomena such as ethnic cleansing and genocide have also appeared in historical and less developed societies. In less developed societies, the most important non-labor factor of production is land, while in industrialized societies the most important non-labor factor of production is capital.[3] It seems likely that the elasticity of substitution between labor and capital is higher than the elasticity of substitution between labor and land, explaining why discrimination as domination is more likely to appear in industrialized societies.

One major exception to the generalization that discrimination as domination is more likely to appear in modern, industrialized societies is discrimination against Jews in medieval and early modern Europe. Jews in this period were subject to occupational restrictions, similar to the restrictions placed on Blacks in apartheid South Africa and the Jim Crow South, but Jews were not completely excluded from the labor market (Becker and Pascali, 2019).[4] Discrimination against Jews is consistent with the argument that discrimination as domination is more likely to appear when assumption 1 holds, that is, when labor and non-labor factors of production are not too complementary. As discussed by Becker and Pascali (2019), Jews in this period provided *skilled* labor, competing with other providers of skilled labor such as monks and members of city guilds. Unlike unskilled labor, skilled labor was not complementary to land during this period. Thus, occupational discrimination against Jews is consistent with my model.

2.3.3 Effects of changing group sizes on wages under discrimination

Next I show how changing the size of the oppressed group affects the dominant group wage under optimal discrimination:

Proposition 3. *Suppose that assumption 1 holds. Then the dominant group wage w_d is increasing in the size of the oppressed group α_o.*

Proof. See appendix. □

Proposition 3 characterizes the effect of a change in the size of the oppressed group on the wage structure under discrimination. As the oppressed group becomes larger, the dominant group wage increases. Intuitively, discrimination redistributes income away from the oppressed group and towards the dominant group, and the larger the oppressed group, the more oppressed group income is available to redistribute. An implication of this result is that the dominant group may be willing to expend resources to increase the size of the oppressed group, for example by promoting immigration (or preventing emigration) by members of the oppressed group.

[3]For example, Esteban et al. (2015) argue that the Rwandan genocide was motivated by conflicts over access to land.

[4]While Becker and Pascali (2019) emphasize the role of occupational restrictions on *Christians* in generating occupational specialization, they also note that there were occupational restrictions on Jews. Consider for example the following quote: "The city guilds forced the Jews out of the trades and the regular channels of commerce; this coincided with the stricter appliance of the church ban on usury in the 12th to 13th centuries. The combination of circumstances made moneylending and pawnbroking the main occupation of Jews in Germany" (Becker and Pascali 2019, online appendix B, quoting the Encyclopedia Judaica).

Like proposition 1, proposition 3 does not hold if the set of reserved tasks is exogenously fixed. If the set of reserved tasks is exogenously fixed, then for sufficiently large σ, the dominant group wage w_d is strictly decreasing in the size of the oppressed group α_o. Proposition 4 presents this result formally:

Proposition 4. *Suppose that σ and R are exogenously fixed. For any $R < 1$, there exists σ sufficiently large that w_d is strictly decreasing in α_o.*

Proof. See appendix. □

3 Applying the model to apartheid

In this section I apply the results discussed above to South African apartheid. The beginning of the apartheid regime in South Africa is usually dated to the victory of the National Party in the (racially segregated) election of 1948, which began a continuous period of National Party rule until the transition to full democracy in 1994. While the National Party ran in 1948 on a promise of increased discrimination against non-Whites and especially against Blacks, the details of how to deliver on this promise were left open. Thus, throughout the early years of the apartheid era, there was significant debate within the National Party about exactly how the new racial order would be organized. There were two main factions within the National Party, supporting two quite different political programs.[5] The first program was known as "total apartheid".[6] Proponents of total apartheid proposed to expel Blacks from White areas of South Africa, including South Africa's cities, the best agricultural areas, and the areas containing the largest mineral deposits, and to split the territory of South Africa into separate, independent, racially homogeneous states for Blacks and Whites. Had it been implemented, this program would have completely removed Black workers from the White economy. The total apartheid program can be explained as an instance of either taste-based or factor-based institutionalized discrimination. Total apartheid corresponded to taste-based discrimination to the extent to which the goal of total apartheid was to reduce the frequency of interactions between Blacks and Whites. Total apartheid corresponded to factor-based institutionalized discrimination to the extent to which the goal of total apartheid was to exclude Blacks from access to non-labor factors of production such as high-quality farmland, underground mineral deposits, and physical capital located in South African cities.

While total apartheid was supported by a significant faction of the National Party, the larger faction supported a different program referred to as "practical apartheid", or in Afrikaans as "baasskap", which translates literally as "boss-ship" or "dominance". The baasskap faction included the first two prime ministers of apartheid South Africa, D. F. Malan and J. G. Strijdom, and so for the most part the baasskap

[5]Posel (1987, 1991) and Kuperus (1999) discuss the debates between National Party factions in the early years of apartheid.

[6]The Afrikaans word "apartheid" translates as "apartness" or "separation" and so total apartheid means "total separation" in English.

program, and not the total apartheid program, was enacted into policy.[7] Proponents of baasskap accepted and supported the continuing growth of the Black population in South African cities and other White areas. The key for proponents of baasskap was not that Blacks should be removed completely from the White economy, but rather that Blacks should participate in the White economy under conditions of inequality. Thus Kuperus (1999, p. 86) describes the views of the first apartheid prime minister, D. F. Malan, as follows: "[Apartheid] did not entail the total separation of races into political, economic, and social arenas; instead Malan 'envisioned local segregation in which inequality would be firmly maintained in all interracial dealings"'. In fact, proponents of baasskap believed that continued Black participation in the White economy was necessary to ensure White prosperity. According to Posel (1991, p. 133), the baasskap faction believed that "White political and economic supremacy presupposed a stable and flourishing economy, built on the back of a predominantly African workforce." Not only did the baasskap faction not support expulsion of Blacks from White areas, but many policies associated with the baasskap faction were explicitly designed to increase the amount of Black participation in the formal White economy. For example, South African tax and land use policy was explicitly designed to force Blacks to seek formal employment in the White economy by forcing Blacks to acquire currency and by making traditional forms of herding and subsistence agriculture infeasible (Feinstein, 2005; Gwaindepi and Siebrits, 2020).

The signature policy of the baasskap program was known as "job reservation" or the "colour bar". Under job reservation, a wide variety of jobs were reserved for Whites. In some cases, job reservations were explicitly enforced by law, while in other cases job reservations were enacted *de facto* through minimum wage regulations, union membership requirements, and training requirements that were in practice impossible for non-Whites to satisfy. While job reservation had existed in South Africa prior to the apartheid era, the apartheid government greatly expanded the scope of job reservation and extended job reservations to nearly all sectors of the economy. Feinstein (2005) discusses job reservation in South Africa prior to the introduction of apartheid and the expansion of job reservation under apartheid. Largely as a result of job reservation, Black wages were dramatically lower than White wages. For example, Feinstein (2005, p. 134) finds that White wages were 10 times higher than Black wages in the South African mining industry in 1935 (when job reservations had been imposed in mining but not yet in all other sectors of the economy). This wage ratio seems much larger than can be explained by differences in human capital between Blacks and Whites. For example, the ratio of unskilled to skilled male wages in the United Kingdom in the same year was 0.69.

[7]The third apartheid prime minister, Hendrik Verwoerd, was more sympathetic to the total apartheid program and attempted to enact some aspects of this program into policy. In particular, Verwoerd created the "homelands", nominally independent states for Blacks. However, the large majority of the putative citizens of each homeland continued to work (and often reside) outside of their homelands, either as migrant workers in urban areas or in White-owned farms or mines. The creation of the homelands thus largely failed to create truly separate economies for members of different racial groups. After Verwoerd the South African state became preoccupied with responding to various external and internal threats, and few new policies from either the baasskap or the total apartheid programs were enacted.

The baasskap program corresponds closely to my model of discrimination as domination. Under baasskap, Blacks continued to participate in the White economy, and indeed Whites pursued policies to increase the number of Blacks participating in the White economy. However, job reservation forced Blacks into an economic role that was constrained and inferior relative to the role played by Whites. Job reservation increased White wages and reduced Black wages, benefitting White workers at the expense of Black workers.

4 Additional results

In this section I discuss some additional implications of my model and apply these implications to apartheid South Africa.

4.1 Socially heterogeneous firms

Section 2.1 argues that in the free market, the allocation of workers from different social groups to tasks is indeterminate. In particular, if there are multiple firms, it is consistent with the free market that all the tasks within each firm are performed by members of the same social group, so that all firms are socially homogeneous. In contrast, if the elasticity of substitution between tasks within firms is lower than the elasticity of substitution between tasks in different firms, then proposition 1 implies that under optimal discrimination there are both reserved and unreserved tasks within each firm. In this case, all firms are socially heterogeneous under discrimination. Thus my model implies that optimal discrimination can increase the prevalence of socially heterogeneous firms relative to the free market. This result is the opposite of the argument in Becker (1957) that discrimination decreases the prevalence of socially heterogeneous firms.

A good example of this phenomenon comes from the mining industry, which was the most important industry in South Africa during the apartheid era. Prior to the introduction of job reservation, the consensus view in the mining industry was that underground mining jobs would soon be held exclusively or nearly exclusively by Blacks, as Black labor was cheaper than White labor. A report by the Mining Regulation Commission in 1925 quoted the view of the Government Mining Engineer, as follows: "I have no reason to doubt that, as natives become more skilled in various occupations, economic law will in years to come operate as it always has, and that the more expensive white man will be replaced to an increasing degree by native labour.... The temptation to the employer to put [Black workers] in the place of the more expensive white man becomes irresistible" (Feinstein (2005), p. 88). One reason that Black workers were predicted to replace White workers was the Black workers were frequently better at mining jobs than White workers, even in the skilled jobs that had historically been performed by Whites. For example, a 1907 government inquiry into the mining industry reports one manager discussing Black workers as follows: "We have some of the [Black workers] who are better machine-men than some of the white men.... Can they place holes [for blasting]? - Yes they can place the holes, fix up the machine, and do everything that a white man can do, but, of course, we are not allowed to let them blast" (Feinstein (2005), p. 88). Notably, mining industry executives believed that Black workers were suitable even for jobs with substantial supervisory responsibilities. The 1925 Mining Regulatory Commission Report states that, "Taking general mining as skilled work, as it surely it, there is an abundance of examples of what are virtually encroachments of the native into it.... [This] has led to the employment of a large number of [Black workers] in what is essentially a skilled position, where they are called upon to exercise over their subordinates wide powers of control and supervision" (Feinstein (2005), p. 88). Legally enforced job reservations were imposed in the mining industry in response to the perception that underground mining jobs would soon be monopolized by Blacks. As Feinstein (2005), p. 88, puts it, "It was clear [to the Mining Regulations Commission] that Africans must be prevented from performing such work, not because they lacked the competence to do it but, on the contrary, precisely because they were, or soon would be, competent. A new colour bar act was urgently required."

As a result of job reservations for Whites, significant numbers of Whites and Blacks continued to work together underground in the mines throughout the apartheid era. Underground mining work is dangerous and requires workers to trust each other and work together in uncomfortable and extremely tightly enclosed conditions. These are exactly the kinds of interracial interactions that people with a Beckerian taste for discrimination would like to avoid, but apartheid regulations increased the frequency of these kinds of interracial interactions. This result is difficult to explain in a model of Beckerian taste-based discrimination, but it is consistent with my model.

4.2 The return to capital under discrimination

Consider the following proposition:

Proposition 5. *In the limit as $\underline{\sigma}$ approaches 0,*

1. *Aggregate labor supply approaches $L = \alpha_d + \alpha_o$, which is the same as aggregate labor supply in the free market.*

2. *The return to the non-labor factor of production Z approaches the return in the free market.*

Proof. See appendix. □

Proposition 5 shows that for $\underline{\sigma}$ approaches 0, the return to the non-labor factor of production under discrimination approaches the return to the non-labor factor in the free market. Of course, in reality it is unlikely that $\underline{\sigma} = 0$, and if $\underline{\sigma} > 0$ then the return to the non-labor factor is lower under optimal discrimination than in the free market. However, proposition 5 can be interpreted as showing that the negative effect of discrimination on the return to the non-labor factor may be relatively small. Intuitively, discrimination has two effects on the return to the non-labor factor of production. First, by increasing the wage for reserved tasks, discrimination reduces the return to the non-labor factor. Second, by reducing the wage for unreserved tasks, discrimination increases the return to the non-labor factor. The second effect partially offsets the first effect, reducing the overall negative effect of discrimination on the return to the non-labor factor.

This result sheds new light on a major academic debate about the causes of apartheid known as the liberal-radical debate. Liberals such as Hutt (1964), Horwitz (1967), and Lipton (1985), heavily influenced by Becker (1957), argued that apartheid reduced the return to capital by driving up the cost of labor, and that capital owners therefore formed an anti-apartheid political constituency. Radicals such as Johnstone (1970), Trapido (1971), Wolpe (1972), and Legassick (1974), drawing on Marxist traditions, argued that apartheid increased the return to capital by reducing the cost of Black labor, and that capital owners therefore formed a pro-apartheid political constituency. My result partially vindicates the radical position by showing how apartheid benefitted capital owners by driving down the cost of Black labor even as it harmed capital owners by driving up the cost of White labor. The overall effect of apartheid on capital owners may have been relatively small. This result helps to explain why, contrary to Lipton (1985), capital owners did little to oppose apartheid before the late 1970s, when the threat of revolutionary unrest against the apartheid regime began to increase following the Soweto riots in 1976. As Thompson (2014) writes (p. 206), "Before the late 1970s no powerful economic interest was fundamentally opposed to apartheid.... Though apartheid imposed costs on the different sectors of business, it also benefitted all of them, and although they criticized specific actions of the government, all sectors accommodated apartheid before 1978."

4.3 Public investment in oppressed group human capital

Proposition 3 shows that the dominant group wage increases in the effective labor supply of the oppressed group. One way to increase the effective labor supply of the oppressed group is by investing in the human capital of the oppressed group. Thus, under optimal discrimination the dominant group may be willing to pay to increase the human capital of the oppressed group, even though the dominant group does not value of the welfare of the oppressed group.

Seekings and Nattrass (2005) find that under apartheid there were net fiscal transfers from Whites to Blacks. Whites paid the large majority of taxes, and the state expended positive amounts on Black education, health care, and old age pensions. Total state expenditure on Black health care was actually higher than total state expenditure on White health care, although expenditure on Black health care was lower than expenditure on White health care on a per capita basis and Whites also consumed private health care services. Seekings and Nattrass (2005) present this finding as a puzzle: given that the state did not value Black welfare, why did the state invest in Black human capital? Some possible answers to this question are not related to my model. For example, the state may have invested in Black health care to prevent epidemic diseases that could also have spread to Whites. However, my model provides an additional possible answer, namely that under discrimination, investment in Black human capital directly increased White wages. This argument may help to explain the perhaps surprisingly large quantity of investment in Black human capital under apartheid.

4.4 Is the set of reserved tasks determined by culture?

I have argued that the dominant group chooses the set of reserved tasks in order to maximize the wage of workers in the dominant group. An alternative theory of task reservations is that the set of reserved tasks is determined by culture. This could be the case if members of the dominant group have tastes for discrimination against members of the oppressed group who perform tasks that are considered culturally inappropriate, as suggested by Hurst et al. (2022).

It seems likely that both cultural and more narrowly economic considerations played a role in determining the set of reserved jobs in apartheid South Africa. For example, one job that was reserved for Whites was the job of elevator operator (Hutt, 1964). Another reserved job was the job of arc welder, although spot welders could be members of any race (Mariotti and van Zyl-Hermann, 2014). The fact that elevator operators were required to be White may have reflected cultural anxieties about Black men being alone in enclosed spaces with White women, rather than narrowly economic concerns. On the other hand, there does not seem to be any obvious cultural reason why arc welders but not spot welders were required to be White. The reservation of arc welding for Whites may have been motivated by economic concerns of the kind presented in my model. My model is relevant to understanding institutionalized discrimination as long as economic considerations play some role in determining the set of reserved tasks, even if the set of reserved tasks is also partly determined by culture.

One way to distinguish task reservations determined by culture from task reservations determined by economics is that culture changes slowly, while economic conditions can change quickly. Thus, task reservations determined by culture are unlikely to change quickly, while task reservations determined by economics may change quickly as economic conditions change. In apartheid South Africa, many job reservations changed frequently. For example, various jobs related to “the manufacture of window or door metal surrounds or to the manufacture of ‘Cliscoe’ windows or ‘Airlite’ louvres” were reserved for Whites in 1958, but these reservations were suspended in 1960, 1968, 1970, and 1974, and cancelled in 1977 (Mariotti and van Zyl-Hermann, 2014). The fact that these job reservations were imposed and then withdrawn multiple times in the space of a few years suggests that these job reservations were motivated by economic rather than cultural considerations.

5 Discrimination versus taxation

I have argued that the purpose of discrimination as domination is to redistribute income from the oppressed group to the dominant group by reducing oppressed group wages and increasing dominant group wages. A natural question is why the dominant group does not prefer to redistribute through the seemingly simpler method of imposing lump-sum taxes on members of the oppressed group and giving the proceeds to members of the dominant group. In some cases, redistribution through taxation may be infeasible. For example, in the US South, it is likely that explicitly race-based taxes would have been ruled unconstitutional under the 14th amendment. In apartheid South Africa, however, explicitly race-based taxes were feasible and were in fact imposed. This observation raises the question of why the apartheid state did not carry out all redistribution through taxation.

One possible answer to this question is that tax collection depends to some extent on voluntary compliance by the population that is being taxed (Andreoni et al., 1998). Citizens are more likely to comply voluntarily with tax collection when they believe that they will benefit from the public goods funded by tax revenue (Levi, 1988; Besley, 2020). In apartheid South Africa, public expenditure was planned primarily to benefit Whites and only incidentally if at all to benefit Blacks. Thus it seems likely that Whites but not Blacks were willing to comply voluntarily with tax collection, and that it was therefore feasible for the state to collect significant tax revenue from Whites but not from Blacks. This argument is consistent with the finding in Seekings and Nattrass (2005) that the large majority of taxes under apartheid were collected from Whites. In contrast, the enforcement of discrimination does not require voluntary compliance from the population that is being discriminated against, and so it may be feasible to redistribute through discrimination in cases where significant redistribution through taxation is infeasible.

Although discrimination does not require voluntary compliance from the population that is being discriminated against, in other ways redistribution through discrimination is more costly than redistribution through taxation. Enforcing discrimination requires the state to collect information about the tasks performed by workers from the oppressed group, in order to ensure that workers from the oppressed group are not performing reserved tasks. In contrast, in order to collect taxes the state typically only needs information about workers' incomes. Thus, in order to enforce discrimination, the state needs to collect more information than is necessary to collect taxes. Collecting information about the tasks performed by workers can be very costly, since under discrimination both workers and employers have an incentive to lie about the tasks that workers perform. Lipton (1985) discusses the difficulties faced by the South African apartheid state in enforcing job reservations in the construction industry, in which Blacks were prohibited from performing various tasks such as masonry in legally designated White areas but not in Black areas. She writes (p. 208), "It proved extremely difficult to enforce the bar in this sector. This was partly because of the pool of blacks trained to work in their 'own' areas..., and partly because of the difficulty of policing thousands of building sites." She continues, "Hundreds of building employers were prosecuted for contraventions of the law - but even larger numbers escaped; and the Minister of Labour became increasingly reluctant to take action against the widespread contraventions." In order to enforce apartheid regulations, the apartheid state was forced to build a massive and extremely costly bureaucracy. Thompson (2014) writes (p. 199), "To administer the laws of apartheid, the bureaucracy grew enormously." Similarly, Feinstein (2005) writes (p. 250) "Implementation and enforcement of apartheid represented a massive waste of resources. An economy suffering from shortages of high-level manpower diverted the work of thousands of well-educated men and women to the bureaucracies established to operate influx controls and all the other administrative measures that were imposed on black people in the name of apartheid."

Because the costs associated with enforcing discrimination are unproductive, discriminatory societies are inefficient relative to non-discriminatory societies. This is a novel form of inefficiency, which differs in particular from the forms of inefficiency under extractive institutions in the taxonomy suggested by Acemoglu (2006). Acemoglu (2006) argues that extractive institutions can be inefficient because extractive institutions distort private-sector labor supply or factor allocation. In my model, proposition 5 shows that in the limit as $\underline{\sigma}$ approaches zero, total output approaches the free market level and so discrimination is not necessarily distortionary in the ways suggested by Acemoglu (2006). Nevertheless, if enforcing discrimination requires unproductive social costs, discrimination may still be inefficient even if it does not distort private-sector labor supply or factor allocation. The discussion of the costs of the apartheid bureaucracy above suggests that this form of inefficiency was a first-order contributor to the overall inefficiency of the South African economy under apartheid.

6 Discrimination as domination in other societies

In this section I apply my model to understanding institutionalized discrimination in societies other than apartheid South Africa. While it is harder to find sharp tests of my model in these contexts, I argue that various features of these societies are broadly consistent with my model.

6.1 The US South under Jim Crow

Wright (1986) discusses job reservation in the US South under Jim Crow. He finds that when Black and White workers performed similar jobs, they received similar wages. However, there were many jobs that were effectively barred to Blacks. He writes (p. 185), "Job discrimination in the better-paying positions was far more important than wage differentials for the same job. Blacks could get the going wage in the unskilled market, but there was a virtual upper limit to their possible progress above that level." Job discrimination was particularly notable in positions which required skills learned on the job. White workers could expect to achieve these positions after several years of experience, but Black workers could not. Wright writes (p. 185), "In Birmingham, for example, with one of the largest concentrations of black industrial labor, more than 80 percent of black workers who stayed for ten years made no upward progress at all toward better jobs during this time. By contrast, half of white workers moved up after ten years." The fact that even experienced Black workers could not advance suggests that Black exclusion from higher paying jobs was not due to lack of skill.

Wright points out that unlike in apartheid South Africa, job discrimination in the US South was for the most part not enforced by law. He suggests that statistical discrimination may explain the observed disparities between Blacks and Whites. I argue, however, that at least part of the observed job discrimination in the South was enforced by informal institutions through the threat of racial violence. Consider, for example, the practice of "whitecapping", which was prevalent throughout the South in the early part of the Jim Crow era. In the 1972 Mississippi statute banning the practice, whitecapping is defined as "threats, direct or implied, of injury to the person or property of another, to intimidate such a person into an abandonment or change of home or employment." Whitecapping threats were often directed at Black workers who took jobs that had customarily been held by Whites. An example of whitecapping comes from the Supreme Court case Hodges v. United States, decided in 1906. The case concerned a group of Whites who had threatened violence to force Black workmen to leave their jobs in a lumber mill in Poinsett County, Arkansas. This kind of violence enacted a form of job reservation, since Poinsett County had a significant Black population who were able to work in other jobs free from violence. The court ruled that the federal government did not have jurisdiction to enforce laws against whitecapping, effectively legalizing whitecapping throughout the South, since all-White Southern state juries were very unlikely to convict Whites of crimes against Blacks. Whitecapping declined in the later part of the Jim Crow era. However, other forms of racial violence developed to threaten Blacks who took jobs that had customarily been reserved for Whites, and employers who offered those jobs. For example, in 1943 the Alabama Dry Dock and Shipping Company (ADDSCO) promoted twelve Black workers to the position of welder, a job that previously had been held exclusively by

Whites, although there were nearly 7,000 Black workers in other jobs within ADDSCO. The next day 4,000 Whites rioted throughout the dockyard, causing 50 injuries and requiring army troops to quell the violence (Nelson, 1993).

6.2 Saudi Arabia

The private sector economy in Saudi Arabia is dominated by non-citizens, mainly consisting of migrants from other Arab countries and from South and South-East Asia. In 2011 non-citizens made up 90% of the private non-oil sector workforce (Peck, 2017). Prior to 2011, the Saudi state made some efforts to encourage citizen employment in the private sector, but these efforts were largely ineffective. This changed in 2011 when the state initiated the Nitaqat program, which imposed quotas limiting the number of non-citizens that firms could hire, combined with harsh penalties if the quotas were not met. Nitaqat effectively forced Saudi firms to discriminate in favor of citizens and against non-citizens. Nitaqat quotas varied by industry, so that tasks associated with low-quota industries were more likely to be effectively reserved for citizens than tasks associated with high-quota industries. Peck (2017) shows that Nitaqat substantially increased private-sector Saudi employment, at the cost of increasing the exit rates of affected firms and reducing total employment in those firms. Notably, Nitaqat did not reduce the total number of migrant workers in Saudi Arabia, although it did shift the distribution of migrant workers across firms (Thiollet, 2022). Nitaqat thus seems to have imposed a form of discrimination as domination on the Saudi private sector, with citizens as the dominant group and non-citizens as the oppressed group.

The case of Saudi Arabia also illustrates my argument that enforcing discrimination is very costly. Thiollet (2022) discusses the difficulty of enforcing Nitaqat regulations, and links the imposition of Nitaqat to investment in increasing Saudi state capacity. Prior to the introduction of Nitaqat, firms were allowed to sponsor migrants directly, with limited regulation by the state. As a result, there existed a substantial shadow migrant labor market, which the Nitaqat reforms attempted to eradicate. Thiollet writes, quoting interviews with Saudi officials (p. 1657), "While 'for a long time, the state has agreed to look away from the shadow economy,' the reforms broke a habit of tolerance towards informal practices." Increasing regulation of the migrant labor market required investment in bureaucratic capacity, such as the creation of new IT systems, the introduction of biometric identification for migrants, and the creation of a new bureaucratic agency, the Vision Realization Office, whose duties included coordination of various migrant labor market regulations. Thiollet quotes an interviewee describing the effects of these efforts, saying (p. 1658) "In 2012 nobody was ready for the reforms. The Ministry of the Interior had no computerized programs. Now all the data are with the Ministry of the Interior." According to Thiollet, the adoption of the new policies "generated intense administrative work" (p. 1659). Thiollet summarizes the effect of Nitaqat as dramatically increasing the capacity of the Saudi state. She writes (p. 1659), "The reforms... entail powerful social engineering that relies upon the administrative ordering and policing. They equip the state with increased structural authority upon market institutions and upon the Saudi society by controlling not only immigrants but also Saudis."

6.3 Policies related to illegal immigrants

In contemporary developed economies, the clearest application of my model is to explaining policies related to illegal immigrants. Immigration laws effectively make it impossible for illegal immigrants to work except in a small number of jobs such as agricultural labor and domestic service, thus imposing a form of discrimination as domination.

Many observers have noted that governments do not seem to do as much as they could to deter illegal immigration. For example, Chiswick (1988), p. 114, writes that "the policy instruments most likely to deter illegal immigration have been ignored." He argues in particular that more is spent on ineffective border enforcement and less on more effective interior enforcement than would be optimal if the goal were to minimize the number of illegal immigrants at a given cost. My model helps to explain this phenomenon. Under optimal discrimination by citizens against illegal immigrants, citizens benefit from the presence of illegal immigrants and so are unlikely to support measures to reduce the illegal immigrant population significantly.

7 Conclusion

In this paper I develop a new theory of institutionalized discrimination, in which the purpose of discrimination is to create a social order in which members of different social groups fill different, hierarchically ranked economic roles. I develop a model in which there are a number of tasks, and in which institutions can reserve some subset of tasks for members of the politically dominant social group. I allow the dominant social group to choose the set of reserved tasks optimally, and I characterize the optimal set of reserved tasks. The dominant social group optimally chooses the set of reserved tasks so that the elasticity of substitution between reserved and unreserved tasks is as low as possible. Opportunities for members of the oppressed group are constrained under discrimination, but the oppressed group has a role to play in society is not completely excluded from the labor market. Under optimal discrimination, the wage of the dominant group is increasing in the size of the oppressed group. I apply my model to understanding discrimination in apartheid South Africa and other discriminatory societies.

The broadest conclusion of my paper is that discrimination results from collective decisions and politics.

This conclusion differs from the main existing theories of discrimination, according to which discrimination results from individual decisions, driven by individual preferences or beliefs. I believe that understanding the institutional and political roots of discrimination is necessary for understanding the most important historical episodes of discrimination, and the persistent effects of these historical episodes in the present.

References

Acemoglu, D. (2006). A simple model of inefficient institutions. *Scandinavian Journal of Economics 108*(4), 515–546.

Acemoglu, D. and D. Autor (2011). Skills, tasks and technologies: Implications for employment and earnings. *Handbook of Labor Economics 4b*, 1043–1171.

Andreoni, J., B. Erard, and J. Feinstein (1998). Tax compliance. *Journal of Economic Literature 36*(2), 818–860.

Beckenbach, E. F. and R. Bellman (1961). *Inequalities*. Berlin: Springer Verlag.

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

Becker, S. O. and L. Pascali (2019). Religion, division of labor, and conflict: Anti-semitism in germany over 600 years. *American Economic Review 109*(5), 1764–1804.

Bergmann, B. R. (1971). The effect on white incomes of discrimination in employment. *Journal of Political Economy 79*(2), 294–313.

Besley, T. (2020). State capacity, reciprocity, and the social contract. *Econometrica 88*(4), 1307–1335.

Chelwa, G., D. Hamilton, and J. Stewart (2022). Stratification economics: Core constructs and policy implications. *Journal of Economic Literature 60*(2).

Chiswick, B. R. (1988). Illegal immigration and immigration control. *Journal of Economic Perspectives 2*(3), 101–115.

Darrity, Jr, W. (2005). Stratification economics: The role of intergroup inequality. *Journal of Economics and Finance 29*(2).

Darrity, Jr, W. (2022). Position and possessions: Stratification economics and intergroup inequality. *Journal of Economic Literature 60*(2).

de La Grandville, O. (1989). In quest of the slutsky diamond. *American Economic Review 79*(3), 468–481.

Esteban, J., M. Morelli, and D. Rohner (2015). Strategic mass killings. *Journal of Political Economy 123*(5), 1038–1086.

Fang, H. and P. Norman (2006). Government-mandated discriminatory policies: Theory and evidence. *International Economic Review 47*(2), 361–389.

Feinstein, C. H. (2005). *An Economic History of South Africa.* Cambridge: Cambdridge University Press.

Gwaindepi, A. and K. Siebrits (2020). 'hit your man where you can': Taxation strategies in the face of resistance at the british cape colony, c.1820 to 1910. *Economic History of Developing Regions 35*(3), 171–194.

Horwitz, R. (1967). *The Political Economy of South Africa.* New York: Frederick A. Praeger.

Hurst, E., Y. Rubinstein, and K. Shimizu (2022). Task-based discrimination. *working paper*.

Hutt, W. H. (1964). *The Economics of the Colour Bar.* London: Institute of Economic Affairs.

Johnstone, F. A. (1970). White prosperity and white supremacy in south africa today. *African Affairs 69*(275), 124–140.

Kamien, M. I. and N. L. Schwartz (1968). Optimal 'induced' technical change. *Econometrica 36*(1), 1–17.

Klump, R. and O. de La Grandville (2000). Economic growth and the elasticity of substitution: Two theorems and some suggestions. *American Economic Review 90*(1), 282–291.

Klump, R., P. McAdam, and A. Willman (2012). The normalized ces production function: Theory and empirics. *Journal of Economic Surveys 26*(5), 769–799.

Krueger, A. O. (1963). The economics of discrimination. *Journal of Political Economy 71*(5), 481–486.

Kuperus, T. (1999). *State, Civil Society and Apartheid in South Africa.* London: Palgrave Macmillan.

Legassick, M. (1974). Legislation, ideology, and economy in post-1948 south africa. *Journal of Southern African Studies 1*(1), 5–35.

Levi, M. (1988). *Of Rule and Revenue.* Berkeley: University of California Press.

Lipton, M. (1985). *Capitalism and Apartheid: South Africa, 1910-1986.* London: Gower Publishing Company.

Lundahl, M. (1982). The rationale of apartheid. *American Economic Review 72*(5), 1169–1179.

Mariotti, M. and D. van Zyl-Hermann (2014). Policy, practice and perception: Reconsidering the efficacy and meaning of statutory job reservation in south africa, 1956–1979. *Economic History of Developing Regions 29*(2), 197–233.

Moro, A. and P. Norman (2004). A general equilibrium model of statistical discrimination. *Journal of Economic Theory 114*, 1–30.

Nelson, B. (1993). Organized labor and the struggle for black equality in mobile during world war ii. *Journal of American History 80*(3).

Norman, P. (2003). Statistical discrimination and efficiency. *Review of Economic Studies 70*(3), 615–627.

Peck, J. R. (2017). Can hiring quotas work? the effect of the nitaqat program on the saudi private sector. *American Economic Journal: Economic Policy 9*(2).

Porter, R. C. (1978). A model of the southern african-type economy. *American Economic Review 68*(5), 743–755.

Posel, D. (1987). The meaning of apartheid before 1948: Conflicting interests and forces within the afrikaner nationalist alliance. *Journal of Southern African Studies 14*(1).

Posel, D. (1991). *The Making of Apartheid, 1948-1961.* Oxford: Clarendon Press.

Seekings, J. and N. Nattrass (2005). *Class, Race, and Inequality in South Africa.* New Haven: Yale University Press.

Small, M. L. and D. Pager (2020). Sociological perspectives on racial discrimination. *Journal of Economic Perspectives 34*(2).

Thiollet, H. (2022). Migrants and monarchs: regime survival, state transformation and migration politics in saudi arabia. *Third World Quarterly 43*(7).

Thompson, L. (2014). *A History of South Africa, 4th Edition.* New Haven: Yale University Press.

Trapido, S. (1971). South africa in a comparative study of industrialization. *Journal of Development Studies 7*(3), 309–320.

Wolpe, H. (1972). Capitalism and cheap labour-power in south africa. *Economy and Society 1*(4), 425–456.

Wright, G. (1986). *Old South, New South: Revolutions in the Southern Economy since the Civil War.* Baton Rouge: Louisiana State University Press.

A Completion of the proof of proposition 1

To show that $R > \alpha_d/(\alpha_d + \alpha_o)$, it suffices to note that if $\underline{\sigma} > 0$, then $w_d = \partial F/\partial L$ if $R \leq \alpha_d/(\alpha_d + \alpha_o)$ and also if $R = 1$. Moreover, if $\underline{\sigma} < \infty$, then L is strictly lower when $R = 1$ than when $R \leq \alpha_d/(\alpha_d + \alpha_o)$, so w_d is greater when $R = 1$ than when $R \leq \alpha_d/(\alpha_d + \alpha_o)$. Thus $R \leq \alpha_d/(\alpha_d + \alpha_o)$ is not optimal.

To show that $R < 1$, I begin by deriving an expression for $\partial L/\partial R$ by differentiating (10) with respect to R:

$$\frac{\partial L}{\partial R} = \frac{1}{\sigma - 1}\left[R^{1/\sigma}\alpha_d^{(\sigma-1)/\sigma} + (1-R)^{1/\sigma}\alpha_o^{(\sigma-1)/\sigma}\right]^{1/(\sigma-1)}\left[\left(\frac{\alpha_d}{R}\right)^{(\sigma-1)/\sigma} - \left(\frac{\alpha_o}{1-R}\right)^{(\sigma-1)/\sigma}\right] \tag{32}$$

By the first part of proposition 1, at the optimum $\sigma = \underline{\sigma}$. If $\underline{\sigma} < 1$, then we have

$$\left.\frac{\partial L}{\partial R}\right|_{R=1} = \frac{1}{\underline{\sigma} - 1}\alpha_d \tag{33}$$

If $\underline{\sigma} > 1$, then we have

$$\lim_{R\to 1}\frac{\partial L}{\partial R} = -\infty \tag{34}$$

Next, differentiating (13) with respect to R, and again using the result that $\sigma = \underline{\sigma}$ from the first part of proposition 1, I derive the following expression:

$$\frac{\partial w_d}{\partial R} = \frac{\partial^2 F}{\partial L^2}\frac{\partial L}{\partial R}\left(L\frac{R}{\alpha_d}\right)^{1/\underline{\sigma}} + \frac{\partial F}{\partial L}\frac{1}{\underline{\sigma}}\left(L\frac{R}{\alpha_d}\right)^{(1-\underline{\sigma})/\underline{\sigma}}\left(\frac{\partial L}{\partial R}\frac{R}{\alpha_d} + L\frac{1}{\alpha_d}\right) \tag{35}$$

If $\underline{\sigma} < 1$, then plugging (33) into (35) yields:

$$\left.\frac{\partial w_d}{\partial R}\right|_{R=1} = \frac{1}{\underline{\sigma} - 1}\left(\frac{\partial^2 F}{\partial L^2}L + \frac{\partial F}{\partial L}\right) \tag{36}$$

Assumption 1 implies that (36) is negative.

If $\underline{\sigma} > 1$, then taking the limit of (32) and plugging in $\sigma = \underline{\sigma}$ yields:

$$\lim_{R\to 1}\frac{\partial w_d}{\partial R} = \left(\lim_{R\to 1}\frac{\partial L}{\partial R}\right)\left(\frac{\partial^2 F}{\partial L^2}L + \frac{1}{\underline{\sigma}}\frac{\partial F}{\partial L}\right) \tag{37}$$

Assumption 1 and (34) imply that there exists $\bar{\sigma} > 1$ such that for $\underline{\sigma} < \bar{\sigma}$, (37) is equal to $-\infty$.

Thus there exists $\bar{\sigma} > 1$ such that for all $\underline{\sigma} < \bar{\sigma}$, $\lim_{R\to 1}\partial w_d/\partial R < 0$, which implies that the optimal value of R is less than 1.

B Completion of the proof of proposition 2

If $R \leq \alpha_d/(\alpha_d+\alpha_o)$, then the discrimination constraint does not bind and $\partial L/\partial R = 0$. If $R > \alpha_d/(\alpha_d+\alpha_o)$ (32) gives an expression for $\partial L/\partial R$.

In (32), if $\sigma < 1$ then $1/(\sigma-1) < 0$ and

$$\left(\frac{\alpha_d}{R}\right)^{(\sigma-1)/\sigma} - \left(\frac{\alpha_o}{1-R}\right)^{(\sigma-1)/\sigma} > 0 \tag{38}$$

If $\sigma > 1$ then $1/(\sigma-1) > 0$ and

$$\left(\frac{\alpha_d}{R}\right)^{(\sigma-1)/\sigma} - \left(\frac{\alpha_o}{1-R}\right)^{(\sigma-1)/\sigma} < 0 \tag{39}$$

So in all cases $\partial L/\partial R \leq 0$.

C Proof of proposition 3

From proposition 2, at the optimum, $\sigma = \underline{\sigma}$. Choose some $\bar{\alpha}_d$, $\bar{\alpha}_o$, and $\bar{R}$, and suppose that $(\partial/\partial\alpha_o)w_d(\bar{\alpha}_d, \bar{\alpha}_o, \bar{R}, \underline{\sigma}) < 0$. Choose some $\hat{\alpha}_o < \bar{\alpha}_o$ such that $w_d(\bar{\alpha}_d, \hat{\alpha}_o, \bar{R}, \underline{\sigma}) > w_d(\bar{\alpha}_d, \bar{\alpha}_o, \bar{R}, \underline{\sigma})$. Notice that $L(\bar{\alpha}_d, \hat{\alpha}_o, \bar{R}, \underline{\sigma}) < L(\bar{\alpha}_d, \bar{\alpha}_o, \bar{R}, \underline{\sigma})$ since L is strictly increasing in α_o. Now choose $\hat{R} > \bar{R}$ such that $L(\bar{\alpha}_d, \bar{\alpha}_o, \hat{R}, \underline{\sigma}) = L(\bar{\alpha}_d, \hat{\alpha}_o, \bar{R}, \underline{\sigma})$; $\hat{R}$ exists because L is decreasing in R, with $lim_{R\to 1} L = \lim_{\alpha_o \to 0} L$. Since $\hat{R} > \bar{R}$ and $L(\bar{\alpha}_d, \bar{\alpha}_o, \hat{R}, \underline{\sigma}) = L(\bar{\alpha}_d, \hat{\alpha}_o, \bar{R}, \underline{\sigma})$, inspection of (13) shows that

$$w_d(\bar{\alpha}_d, \bar{\alpha}_o, \hat{R}, \underline{\sigma}) > w_d(\bar{\alpha}_d, \hat{\alpha}_o, \bar{R}, \underline{\sigma}) \tag{40}$$

$$> w_d(\bar{\alpha}_d, \bar{\alpha}_o, \bar{R}, \underline{\sigma}) \tag{41}$$

So the assumption that $(\partial/\partial\alpha_o)w_d(\bar{\alpha}_d, \bar{\alpha}_o, \bar{R}, \underline{\sigma}) < 0$ implies that $\bar{R}$ is not the optimal R. Therefore, if R^* is the optimal R, $(\partial/\partial\alpha_o)w_d(\bar{\alpha}_d, \bar{\alpha}_o, R^*, \underline{\sigma}) \geq 0$. By the envelope theorem, $(\partial/\partial\alpha_o)w_d(\bar{\alpha}_d, \bar{\alpha}_o, R^*, \underline{\sigma}) = (d/d\alpha_o)w_d(\bar{\alpha}_d, \bar{\alpha}_o, R^*, \underline{\sigma})$, completing the proof that w_d is increasing in α_o.

D Proof of proposition 4

Using (13), we have:

$$\frac{\partial w_d}{\partial \alpha_o} = \frac{\partial L}{\partial \alpha_o}\left[\frac{\partial^2 F}{\partial L^2} + \frac{R}{\sigma\alpha_d}\frac{\partial F}{\partial L}\left(L\frac{R}{\alpha_d}\right)^{(1-\sigma)/\sigma}\right] \tag{42}$$

As σ approaches ∞, the second term in the square brackets on the right hand side of (42) approaches

0. By assumption, $\partial^2 F/\partial L^2 < 0$. Therefore, for sufficiently large σ, the term in the square brackets on the right hand side of (42) is negative. In addition, $\partial L/\partial \alpha_o > 0$ for all $\sigma < \infty$ and $R < 1$. So the right hand side of (42) is negative for sufficiently large σ.

E Proof of proposition 5

In the limit as σ approaches 0, L approaches $\min\{\alpha_d/R, \alpha_o/(1-R)\}$. Choose $\epsilon > 0$, and let $R = \alpha_d/(\alpha_d + \alpha_o) + \epsilon$. Then in the limit as σ approaches 0 the oppressed group wage approaches 0 and the dominant group captures the entire payment to labor. Moreover, L is decreasing in ϵ and so by assumption 1 the total payment to labor and hence the dominant group wage are also decreasing in ϵ. So is optimal to set ϵ arbitrarily close to 0, that is to set R arbitrarily close to $\alpha_d/(\alpha_d + \alpha_o)$. In this case aggregate labor supply is arbitrarily close to $L = \alpha_d + \alpha_o$ and the return to the non-labor factor of production $\partial F/\partial Z$ is arbitrarily close to the free market return.