A theory of discriminatory institutions, with applications to apartheid and to the political economy of migration*

James P. Choy

March 4, 2025

Abstract

Institutions in some societies force employers to discriminate. I develop a theory of institutionalized discrimination. Optimal discrimination endogenously creates categories of "good" and "bad" jobs, and assigns workers from different social groups into these different categories. The relative scarcity of labor determines whether discrimination or free labor markets are optimal. When discrimination is optimal, the dominant group benefits from increasing the supply of oppressed group labor. I apply the model to apartheid South Africa and to the regulation of migrant labor in contemporary economies.

Keywords: Discrimination, migration, tasks
JEL Classification Numbers: J71, P48

*Corresponding author: James P. Choy, University of York. E-mail: james.choy@york.ac.uk. I thank Miguel Leon-Ledesma, Martine Mariotti, Paulo Santos-Monteiro, and various conference and seminar participants for helpful comments.

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.

The standard explanation of discriminatory institutions is that discriminatory institutions are designed to provide material benefits to workers in the dominant group, at the expense of workers in the oppressed group and possibly also at the expense of owners of other factors of production such as land or capital (Krueger, 1963; Porter, 1978; Lundahl, 1982; Mariotti, 2012; Hutt, 1964; Lipton, 1985). Discriminatory institutions achieve this goal by reserving certain jobs for members of the dominant group, increasing the wages of workers in 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.[1] 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?

To answer these questions, I construct a model in which there are three social groups. Two of the groups consist of workers, and I label these groups as "dominant" and "oppressed". The third group consists of owners of a non–labor factor of production. Dominant group workers monopolize political power, excluding both oppressed group workers and owners of the non–labor factor. Workers can

[1] Wright (1986) discusses informal job reservations for Whites in the Jim Crow South. An example of the violent enforcement of these job reservations comes from Nelson (1993), who describes the 1943 riot at the Alabama Dry Dock and Shipping Company (ADDSCO). ADDSCO promoted 12 Black workers to the position of welder, a position that had previously been held exclusively by Whites, although there were more than 7,000 Black workers in other positions within ADDSCO. The next day 4,000 Whites rioted throughout the shipyard, requiring army troops to quell the violence.

choose to work in any of a number of different tasks. Final output is produced from an aggregate of labor applied to different tasks and from the non–labor factor. Dominant group workers can use their political power to impose labor market regulations reserving some subset of the available tasks for members of the dominant group. 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. The dominant group chooses these parameters to maximize the wage of workers in the dominant group. Thus the function of task reservations is to redistribute income towards dominant group workers and away from both oppressed group workers and owners of the non–labor factor. The dominant group redistributes through task reservations, instead of through more efficient methods such as lump–sum taxation, because efficient taxation requires a social contract between the state and its citizens as in Levi (1988) and Besley (2020). In a hierarchical society, there may be a social contract between the dominant group and the state, but there is no social contract between the oppressed group and the state. As a result, efficient taxation of the oppressed group is infeasible and so the dominant group chooses to redistribute through inefficient discrimination instead.

I first characterize the optimal elasticity of substitution between reserved and unreserved tasks. I show that the dominant group optimally sets the elasticity of substitution between reserved and unreserved tasks to be as low as possible, so that oppressed group workers and dominant group workers are complements. This result describes an economy in which members of different social groups perform economically distinct functions. Dominant group workers hold "good" jobs with high wages, while oppressed group workers hold "bad" jobs with low wages. However, what makes different jobs good or bad is not the intrinsic features of different jobs but rather the social institutions that artificially increase the wages of dominant group jobs while suppressing the wages of oppressed group jobs.

Next I characterize the optimal size of the set of reserved tasks. Under discrimination, oppressed group workers affect dominant group workers in two ways. First, oppressed group workers harm dominant group workers by competing with dominant group workers for access to the non-labor factor of production. I refer to this effect as the competition effect. Second, oppressed group workers benefit dominant group workers by providing complementary labor, increasing the productivity of dominant group workers. I refer to this effect as the complementarity effect. When the competition effect is strong relative to the complementarity effect, it is optimal for the dominant group to protect

dominant group workers from competition by reserving many tasks. When the competition effect is weak relative to the complementarity effect, it is optimal for the dominant group to reserve fewer tasks, thereby allowing dominant group workers to benefit more from the complementary labor of members of the oppressed group.[2]

I apply these ideas to argue that the choice between discrimination and free labor markets depends on the relative abundance of labor. When labor is abundant relative to the non-labor factor of production, the competition effect is relatively strong, and so it is optimal for dominant group workers to reserve many tasks. In this case optimal discrimination is severe and the benefits of discrimination for the dominant group relative to the free market are large. When labor is scarce, the competition effect is relatively weak, and so it is optimal to reserve fewer tasks. In this case optimal discrimination is less severe and the benefits of discrimination for the dominant group relative to the free market are small. If there is a fixed cost to imposing discrimination, then the dominant group prefers to impose discrimination when labor is abundant and to allow free labor markets when labor is scarce.

The optimal labor market regime determines the attitude of the dominant group towards increasing the size of the oppressed group. Under free labor markets, the wage of the dominant group is decreasing the size of the oppressed group. In contrast, under discrimination, the wage of the dominant group is increasing in the size of the oppressed group. Thus, under discrimination, 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.

I apply my model to the understand the history of discrimination in South Africa. In the 19th and early 20th centuries, labor market policy in South Africa was essentially libertarian, with few restrictions on jobs that could be performed by members of different racial groups. In the 1920s and 1930s, the South African state began to reserve progressively larger classes of jobs for Whites. This process culminated in election of 1948, which is conventionally regarded as the beginning of the apartheid regime. Discrimination remained very severe in South Africa throughout the 1950s and 1960s. However, in the 1970s job reservations for Whites began to be relaxed, allowing Blacks to perform jobs from which they had previously been barred. Nearly all job reservations were removed by 1984. The end of discrimination in South Africa occurred in the context of continuing White

[2]Krueger (1963) and Porter (1978) only discuss the competition effect. Mariotti (2012) only discusses the complementarity effect. Lundahl (1982) implicitly includes both effects, but Lundahl (1982) does not define these effects explicitly or derive any of the consequences of the two effects discussed below.

political control, as non-Whites did not gain the franchise until 1994.

I argue that changes in the severity of discrimination in South Africa were caused by changes in the relative abundance of labor. In the 1920s, a wave of unskilled migrants from rural areas to cities led to a sharp increase in the abundance of both White and Black unskilled labor. This led to demands from unskilled Whites for protection against competition from unskilled Blacks, and the state responded by imposing job reservations for Whites. By the 1970s, nearly all Whites had acquired at least some skills, so competition from unskilled Blacks was no longer relevant to Whites. Moreover, in the 1970s changes in labor demand led to a severe shortage of both White and Black skilled labor. In this context, job reservations were less beneficial to Whites, and so job reservations were relaxed and ultimately eliminated.

I also show that during the period when the South African state was discriminating against Blacks, the South African state explicitly rejected proposed policies designed to remove Blacks from the White economy, and instead imposed policies that were explicitly designed to increase the supply of Black labor into the White economy. These policies are broadly inconsistent with taste-based theories of discrimination, according to which discrimination is meant to reduce contact between Whites and Blacks. However, these policies are consistent with the prediction of my model that under discrimination, White workers benefit from increasing Black labor supply.

In the contemporary world, the primary application of my model is towards understanding the political economy of migration, with native workers as the dominant group and migrant workers as the oppressed group. I discuss a number of examples of societies in which certain jobs are reserved for native workers. My model shows that under optimal job reservation for native workers, increased immigration benefits natives. Thus, job reservations for native workers can help to build political support for immigration. As birthrates in the developed world decline, migration from developing countries to developed countries is likely to become more important. However, increasing political backlash against migration suggests that large-scale migration with free labor markets is likely to be politically infeasible. I therefore predict that in the future, job reservations for native workers will become more prevalent throughout the developed world. This transition will have important consequences not only for economics, but also for ethics, political philosophy, law, and international relations.

In the literature, the paper technically 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. Hsieh et al. (2019) present a model with task-specific levels of discrimination that can change over time, but they do not explain why these changes over time occur. Mariotti (2012) presents a model in which the set of reserved tasks is partially endogenous, but in which there are only three discrete categories of tasks that can be reserved.

More generally, my model is related to theories of extractive institutions (Acemoglu and Robinson, 2012). Acemoglu (2006) presents a taxonomy of ways in which extractive institutions may be inefficient. Discriminatory institutions are most closely related to what Acemoglu (2006) refers to as the factor price manipulation function of extractive institutions. According to this argument, extractive institutions redistribute income by artificially suppressing the supply and hence raising the price of some factors of production. In Acemoglu (2006), different factors of production are defined exogenously by technology, while in my model social institutions endogenously transform dominant group labor and oppressed group labor into effectively distinct factors of production. Mukand and Rodrik (2020) and McGee (2024) present models of extractive institutions in which society is divided by an indicator of social identity such as race, but they do not describe extraction through labor market discrimination as in my paper.

Finally, my model is related to theories of dual labor markets such as Lewis (1954) and Harris and Todaro (1970). Lewis (1954) and Harris and Todaro (1970) construct models in which the labor market is divided into a sector that is protected by minimum wage legislation and a sector that is not. In my model the two labor market sectors are the sector of reserved jobs and the sector of unreserved jobs.

1.1 Technical contribution

Technically, my model contributes to the theory of constant elasticity of substitution (CES) production functions. The proof of the key lemma in my model uses of the concept of a normalized CES production function, introduced by de La Grandville (1989) and further developed by Klump and de La Grandville (2000). Recent work on normalized CES productions functions is reviewed by Klump et al. (2012). Leon-Ledesma and Satchi (2019) apply related ideas to understanding directed

technical change. (The analogy with my paper is clearer in an earlier version, Leon-Ledesma and Satchi (2011).) Given the close analogy between my work and results in the theory of directed technical change, the proofs of my theorems may be of independent interest.

2 Model setup

2.1 An example economy

I begin with an example of an economy. This example motivates the more general economy described in section 2.2 below.

Consider a society which contains three social groups. Two of the groups consist of workers, which I will label "dominant" and "oppressed". The third group is the group of owners of a non-labor factor of production such as physical capital, human capital, or land. Each group contains a continuum of members. Let the measures of the sets of workers in the dominant and oppressed groups be $\alpha_d \in [0, \infty)$ and $\alpha_o \in [0, \infty)$, respectively. The measure of the set of owners of the non-labor factor is zero, representing the idea that ownership of the non-labor factor is unequally distributed. Dominant group workers monopolize political power, excluding both oppressed group workers and owners of the non-labor factor. Dominant group workers exclude oppressed group workers from power through political suppression, for example by denying oppressed group workers the vote. Dominant group workers can exclude owners of the non-labor factor from political power even if owners of the non-labor factor are able to vote, because the number of owners of the non-labor factor is small.

In the South African context, the division of society into dominant group workers, owners of the non-labor factor, and oppressed group workers can be interpreted as representing the division of society into relatively poor but more numerous Afrikaans–speaking Whites, relatively rich but less numerous English–speaking Whites, and Blacks.[3] Throughout the apartheid era, political power was monopolized by the political party representing Afrikaans–speaking Whites, excluding both Blacks and the party representing English–speaking Whites. I discuss the history of apartheid in more detail below.

Each worker inelastically supplies one unit of labor, and chooses to supply this unit of labor to

[3] "Blacks" in the South African context includes the indigenous peoples of South Africa as well as the mixed–race "Colored" population and Indian immigrants, all of whom were denied full political participation under apartheid.

one out of a number of different tasks. All tasks require the same level of skill. Therefore, any worker is physically able to perform any task.

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. There are a continuum of divisions in the firm and a continuum of tasks within each division, and 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 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 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 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 representative firm is a price taker and makes zero profit. Therefore all workers are paid their marginal products. The marginal product and hence the wage of a worker who supplies labor to task i in division j is $\partial F/\partial \ell_{i,j}$.

Dominant group workers can use their 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. The dominant group can choose to reserve any subset of the set of all tasks indexed by i, j. I discuss the dominant group's objective function in section 2.2 below.

Since in the free market, workers can legally choose to supply labor to any task, 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 $Y = F(Z, \alpha_d+\alpha_o)$, and the wage for workers in both groups is $w = \partial F/\partial L$. 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 (out of the many) possible 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 there are fewer dominant group workers per task than there are oppressed group workers per task, and so the marginal product and hence the wage is higher for reserved tasks than the wage for unreserved tasks. Thus, 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. Thus, each reserved task is performed by the same number of workers, and likewise each unreserved task is performed by the same number of workers. 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 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 $\sigma = \tau_1$. The marginal product and hence the wage of dominant group workers is $w_d = \partial F/\partial \alpha_d = (\partial F/\partial L)(\partial L/\partial \alpha_d)$, and the marginal product and hence the wage of oppressed group workers is $w_o = \partial F/\partial \alpha_o = (\partial F/\partial L)(\partial L/\partial \alpha_o)$.

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 is:

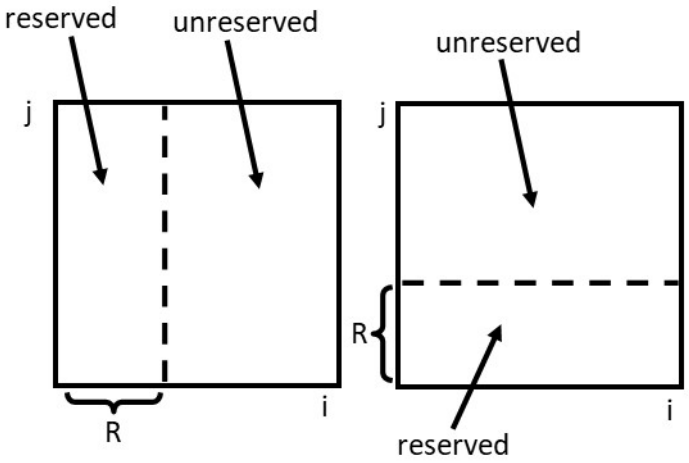
$$L = \left[R \left(\frac{\alpha_d}{R} \right)^{(\tau_2 - 1)/\tau_2} + (1 - R) \left(\frac{\alpha_o}{1 - R} \right)^{(\tau_2 - 1)/\tau_2} \right]^{\tau_2/(\tau_2 - 1)} \tag{8}$$

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

Figure 1 represents graphically the two ways of dividing the set of tasks into reserved and unreserved tasks.

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 independently decide whether the elasticity of substitution between dominant and oppressed group workers in the aggregate production function is $\sigma = \tau_1$ or $\sigma = \tau_2$. As before, the marginal product and hence the wage of dominant group workers is $w_d = \partial F/\partial \alpha_d =$

Figure 1: Reserved and unreserved tasks



This figure shows two possible ways of reserving a measure of tasks R. The space is the space of tasks indexed by i and j, each with measure 1. The figure on the left depicts reserving a measure R of tasks within each division j. The figure on the right depicts reserving all tasks within a measure R of divisions, and leaving all tasks within the remaining divisions unreserved.

$(\partial F/\partial L)(\partial L/\partial \alpha_d)$, and the marginal product and hence the wage of oppressed group workers is $w_o = \partial F/\partial \alpha_o = (\partial F/\partial L)(\partial L/\partial \alpha_o)$.

With finer distinctions between tasks it is possible for the elasticity of substitution between reserved and unreserved tasks to take on additional values. For example, suppose that each task is composed of a continuum of subtasks with measure 1, indexed by a third dimension, k, and suppose that the elasticity of substitution between subtasks is τ_3. Then by reserving a measure R of the subtasks that compose each task, the dominant group could set the elasticity of substitution dominant and oppressed group workers in the aggregate production function equal to τ_3. With sufficiently fine distinctions between tasks the dominant group could choose many different elasticities of substitution for any given value of R.

In principle, it would also be possible for the dominant group to partition the set of tasks in more complicated ways. The general problem of how to optimally partition the set of tasks is an optimization problem with infinite (and indeed, uncountable) dimension, since the dominant group would have to choose for each of an infinite number of tasks whether the task should be reserved or unreserved. I am unable to solve this problem in general. Therefore, in what follows I simplify the

dominant group's problem by assuming that the dominant group chooses the set of reserved tasks only by choosing R and σ.

2.1.1 Application

To fix ideas, it may be helpful to present a concrete example of how it is possible to vary the parameters in my model by choosing the set of reserved tasks appropriately. In the mid-20th century steel industry, a riveting team consisted of four members: the heater, the catcher, the riveter, and the bucker. All four tasks required similar levels of skill, and all four team members were required for production, so within each team, the different tasks were complements. In 1950s Alabama, both Whites and Blacks worked in riveting teams, as discussed in Norrell (1986). One possible way to organize a given number of White and Black workers would have been to have some teams in which all four workers were White and other teams in which all four workers were Black. In this case, White and Black workers would have been substitutes. In fact, though, riveting teams were not organized in this way. In actual riveting teams, the bucker was always Black, while the other three team members were always White. Under this form of organization, White and Black workers were complements. This example shows how it is possible to choose the set of reserved tasks to vary the elasticity of substitution between reserved and unreserved tasks while holding the sizes of the sets of reserved and unreserved tasks fixed.

2.2 General setup

In this section I present my general modelling setup. The general setup is reduced form, in that I proceed directly from an aggregate production function without describing how this aggregate production function is derived from underlying primitives. However, the aggregate production function is motivated by the example in section 2.1 and so the reader should refer to the primitives defined there.

As in section 2.1, society consists of three groups, dominant workers, oppressed workers, and owners of the non-labor factor. The measures of dominant and oppressed group workers are $\alpha_d \in [0, \infty)$ and $\alpha_o \in [0, \infty)$, respectively, and the measure of owners of the non-labor factor is zero. Final output is a function of an aggregate of labor L and some other factor of production Z:

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

Dominant group workers monopolize political power and control the state. Through its control of the state, the dominant group chooses whether to allow free labor markets or to impose institutionalized discrimination. In the free labor market, following the example in section 2.1, $L = \alpha_d + \alpha_o$. As in section 2.1, the wage for both groups in the free market is then

$$w(\alpha_d, \alpha_o) = \frac{\partial F}{\partial L} \tag{10}$$

Under institutionalized discrimination, following the example in section 2.1, L is a CES function of the sizes of the dominant and oppressed groups:

$$L(\alpha_d, \alpha_o, R, \sigma) = \left[R\left(\frac{\alpha_d}{R}\right)^{(\sigma-1)/\sigma} + (1-R)\left(\frac{\alpha_o}{1-R}\right)^{(\sigma-1)/\sigma} \right]^{\sigma/(\sigma-1)} \tag{11}$$

In the example in section 2.1, by choosing set of reserved tasks, the dominant group could choose R and σ in the aggregate labor 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 section 2.1 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:

$$\sigma \geq \underline{\sigma} \tag{12}$$

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

I assume the following:

Assumption 1. *For all* $L \in [\alpha_d, \alpha_d + \alpha_o]$

1. $\partial F/\partial L > 0$
2. $\partial^2 F/\partial L^2 < 0$
3. $$\frac{\partial}{\partial L}\left(\frac{\partial F}{\partial L} L\right) = \frac{\partial^2 F}{\partial L^2} L + \frac{\partial F}{\partial L} > 0 \tag{13}$$

4. Define

$$\xi(R) = 1 + \frac{1}{\partial L/\partial R}\frac{L}{R} \tag{14}$$

Define

$$\zeta(R) = \frac{\partial^2 F}{\partial L^2}L + \frac{1}{\sigma}\frac{\partial F}{\partial L}\xi(R) \tag{15}$$

For all $\sigma > 0$, either $\zeta(R) < 0$ for all R or $\zeta(R)$ is strictly increasing in R.

The most substantive part of assumption 1 is part 3, which 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 L 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, part 3 of assumption 1 may be satisfied if the elasticity of total output with respect to aggregate labor L is sufficiently large. Part 4 of assumption 1 is a technical assumption that ensures that the model has a unique solution. This assumption is somewhat difficult to interpret, but approximately speaking it states that the third derivative $\partial^3 F/\partial L^3$ is small in absolute value relative to the second derivative $\partial^2 F/\partial L^2$. For example, the assumption holds if $\partial^2 F/\partial L^2$ is constant and $\partial^3 F/\partial L^3 = 0$, since $\partial F/\partial L$ and $\xi(R)$ are strictly increasing in R. The assumption also holds if F is the Cobb-Douglas production function. If F is the CES production function, and the elasticity of substitution between Z and L is small, then the assumption may not hold. However, in this case it is likely that part 3 of assumption 1 does not hold either. I discuss the purpose of assumption 1 in more detail below.

As in section 2.1, the wage of workers in the dominant group under discrimination 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 following expression for the dominant group wage is useful:

$$w_d(\alpha_d, \alpha_o, R, \sigma) = \frac{\partial F}{\partial L}\left(L\frac{R}{\alpha_d}\right)^{1/\sigma} \tag{16}$$

In order to enforce discriminatory regulations, the state must check that each person who holds a reserved job is a member of the dominant group. Thus the cost of enforcing discriminatory regulations is proportional to the number of people who hold reserved jobs, which in equilibrium is equal to the number of members of the dominant group. The cost of enforcing discriminatory

regulations is also proportional to the dominant group wage, since a higher dominant group wage implies a greater incentive to evade discriminatory regulations by employing oppressed group workers in dominant group jobs. Thus, the cost of enforcing discriminatory regulations is $cw_d\alpha_d$, where $c < 1$ is some constant.[4] In order to pay this cost, the state can impose lump–sum taxes t_d and t_o on workers from the dominant and oppressed groups, respectively. I allow for taxes to be negative to allow for fiscal redistribution towards members of some group. Taxation of owners of the non-labor factor of production is infeasible, perhaps because the non-labor factor is mobile across international boundaries. Let ω be an indicator equal to 1 if discriminatory institutions are imposed and 0 otherwise. The state's budget constraint is then:

$$\omega c w_d \alpha_d \leq t_d \alpha_d + t_o \alpha_o \tag{17}$$

The state can impose a maximum penalty π for failing to pay taxes. I assume that the state has low capacity and so $\pi = 0$. In addition, following Besley (2020), I assume that workers may not have purely materialistic preferences. Instead, workers from group i may face an additional psychic cost λ_i from failing to pay taxes. Besley (2020) refers to workers who have these preferences as "civic-minded". Besley (2020) argues that civic-mindedness is a form of reciprocity that motivates workers to support the state voluntarily because they believe that the state will look after their interests in return. Workers who do not expect to receive any benefits from the state are not willing to support the state voluntarily. Since the state is controlled by members of the dominant group and serves the interests of the dominant group, I assume that only members of the dominant group are civic-minded, that is, $\lambda_d > 0$ and $\lambda_o = 0$. Thus, the maximum feasible taxes on the dominant and oppressed groups are respectively $\bar{t}_d = \lambda_d$ and $\bar{t}_o = 0$. I assume that $\lambda_d > cw_d$, so that is feasible to collect all of the funds necessary to pay for discrimination by taxing members of the dominant group.

Because the maximum feasible tax rate on members of the oppressed group is $t_o = 0$, it is not feasible for the state to redistribute from the oppressed group to the dominant group through taxation. Thus if discrimination is imposed, then the state sets $t_d = cw_d$; otherwise, the state sets

[4]The key claim here is that the cost of enforcing discriminatory regulations is a constant fraction of the dominant group wage w_d. In the text I present one set of assumptions that motivate this claim. There are also other ways to motivate the claim. For example, perhaps (as was the case in South Africa) discriminatory institutions lead to international sanctions and boycotts that prevent dominant group members from consuming internationally produced goods, reducing dominant group utility by a constant fraction of income.

$t_d = 0$. The assumptions that generate this result are obviously somewhat contrived and are made for simplicity. The more general argument is efficient taxation requires a social contract between the state and its citizens, as argued by Levi (1988) and Besley (2020). A discriminatory state may have a social contract with the dominant group, but there is no social contract between the state and the oppressed group. Thus, efficient taxation of the oppressed group is infeasible, and so if the state wants to redistribute from the oppressed group to the dominant group it must do so through some method other than taxation. Apartheid South Africa was a democracy for Whites, and so it is reasonable to suppose that a social contract existed between the South African state and the White population. However, Blacks were excluded from political participation and there was no social contract between the state and the Black population. The apartheid state did attempt to redistribute through taxation by imposing explicitly race–based taxes. However, taxes on Blacks did not generate very much revenue. Seekings and Nattrass (2005) find that there was actually net fiscal redistribution from Whites to Blacks under apartheid, as the apartheid state spent money on Black education, health care, and old-age pensions. The amounts spent on Blacks were small relative to the amounts spent on Whites in per-capita terms, but were greater than the amount of revenue collected from Blacks through taxation. The observation that tax revenue raised from Blacks was low supports the argument that the apartheid state was not able to achieve as much redistribution as it would have liked through taxation alone.

The dominant group makes two decisions. First, the dominant group decides whether to impose discrimination or not. If the dominant group does not impose discrimination, then dominant group workers receive the free market wage w. If the dominant group does impose discrimination, then dominant group workers receive the discriminatory wage w_d net of the per–capita cost cw_d necessary to pay for discriminatory regulations. The dominant group's objective is to maximize the dominant group wage net of the fixed cost. Thus, the dominant group solves:

$$\max\{w(\alpha_d, \alpha_o), \max_{R,\sigma}(1-c)w_d(\alpha_d, \alpha_o, R, \sigma)\} \tag{18}$$

subject to the constraint (12).

By enforcing discrimination, the dominant group transfers income towards dominant group workers and away from both oppressed group workers and owners of the non-labor factor. In apartheid South Africa, the standard view is that the balance of political power was held by poor (primarily

Afrikaans–speaking) White workers, who imposed discrimination to benefit themselves at the expense of Black workers and also at the expense of (primarily English–speaking) White capital and land owners (Hutt, 1964; Lipton, 1985; Seekings and Nattrass, 2005; Feinstein, 2005; Thompson, 2014). I discuss the specifics of South African politics in more detail in section 3 below.

3 Results

3.1 The optimal elasticity of substitution between reserved and unreserved tasks

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

$$MRTS = \frac{\partial L/\partial \alpha_d}{\partial L/\partial \alpha_o} = \left[\frac{R}{(1-R)}\frac{\alpha_o}{\alpha_d}\right]^{1/\sigma} \tag{19}$$

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 functions 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

$$\left.\frac{\partial L/\partial \alpha_d}{\partial L/\partial \alpha_o}\right|_{\bar{\alpha}_d, \bar{\alpha}_o, R(\bar{\alpha}_d, \bar{\alpha}_o, \bar{\mu}, \sigma)} = \bar{\mu} \tag{20}$$

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

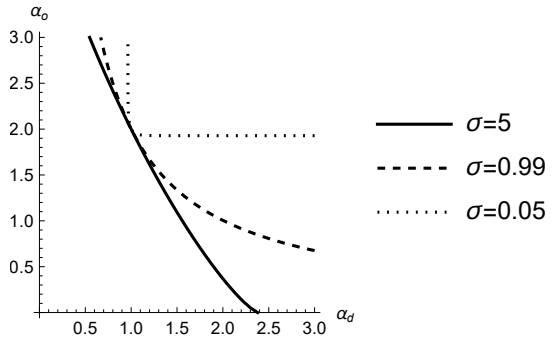
$$\hat{L}(\alpha_d, \alpha_o; R(\bar{\alpha}_d, \bar{\alpha}_o, \bar{\mu}, \sigma), \sigma) = \left[(R(\bar{\alpha}_d, \bar{\alpha_o}, \bar{\mu}, \sigma))^{1/\sigma}\alpha_d^{(\sigma-1)/\sigma} + (1 - R(\bar{\alpha}_d, \bar{\alpha}_o, \bar{\mu}, \sigma))^{1/\sigma}\alpha_o^{(\sigma-1)/\sigma}\right]^{\sigma/(\sigma-1)} \tag{21}$$

Figure 2 shows the isoquants of a family of CES aggregate labor 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*

Figure 2: A family of normalized CES aggregate labor functions



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

in σ*.*

Proof. See appendix. □

Lemma 1 states that when R is varied to hold the wage ratio w_d/w_o fixed, aggregate labor L is strictly decreasing in 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* $\underline{\sigma} < \infty$ *then under optimal discrimination the dominant group wage* w_d *is strictly greater than the free market wage and the oppressed group wage* w_o *is strictly less than the free market wage.*

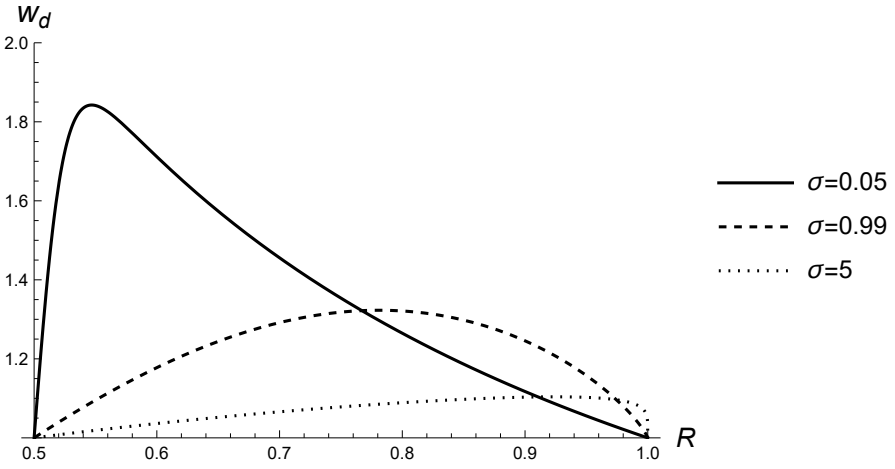
Proof. From lemma 1, decreasing σ while varying R to hold the wage ratio w_d/w_o fixed increases aggregate labor L. Assumption 1 states that increasing aggregate labor L increases the total payment to labor. If the total payment to labor increases while the wage ratio remains fixed, the wage of the

dominant group must increase. Therefore, it is always possible to increase the dominant group wage by reducing σ, so it is optimal to set σ as low as possible.

The second part of proposition 1 follows from the observation the free market wage is equal to the dominant group wage under discrimination with $\sigma = \infty$. If $\underline{\sigma} < \infty$, the first part of proposition 1 then implies that the wage for dominant group workers under optimal discrimination is greater than the free market wage. Since aggregate labor is lower under optimal discrimination than in a free market, and since the total payment to labor is increasing in aggregate labor by assumption 1, the total payment to labor is lower under optimal discrimination than in a free market. Thus the oppressed group wage must be lower under optimal discrimination than in a free market. □

It is important for proposition 1 that both R and 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 $F(Z, L) = L$, $\alpha_d = \alpha_o = 0.5$, for values of R between 0.5 and 1 and for $\sigma = 0.05$, $\sigma = 0.99$, and $\sigma = 5$

3.1.1 Discussion

Proposition 1 describes an economy in which members of different social groups perform economically distinct, complementary tasks. The wages for dominant group tasks are greater than the wages of oppressed group tasks and so members of the dominant group hold "good" jobs while members of the oppressed group hold "bad" jobs. Since all tasks are intrinsically symmetrical in the model, it is not the intrinsic features of different jobs that determine which jobs are good or bad. Instead, social institutions artificially inflate the wage for dominant group jobs while suppressing the wage for oppressed group jobs. In this way, social institutions redistribute income from the oppressed group to the dominant group.

In my model, tasks performed by the oppressed group are bad because they pay lower wages. In reality, the jobs held by members of oppressed social groups are often bad for non–wage reasons. For example, members of oppressed groups are often restricted to jobs that are dangerous, unpleasant, or demeaning. In a free labor market, these "bad" jobs would pay higher wages than similarly skill–intensive "good" jobs in order to compensate for their negative amenities. In a discriminatory labor market, bad jobs may pay wages equal to or lower than the wages of similarly skill–intensive good jobs.

Mariotti (2012) and Hurst et al. (2024) discuss some other dimensions along which jobs may be "bad". Mariotti (2012) argues that highly skill-intensive jobs are often reserved, while less skill–intensive jobs are often unreserved. In a free labor market, the wages of highly skill–intensive jobs and less skill–intensive jobs equalize after accounting for variation in training costs and in innate ability. However, in a discriminatory labor market in which highly skill–intensive jobs are reserved, members of the oppressed group in less skill–intensive jobs may receive lower wages even after accounting for their lower training costs and possible ability differences. In many cases, highly skilled jobs may not be reserved *de jure*, but are reserved *de facto* by regulations making it impossible for members of the oppressed group to acquire the necessary skills. This was the case in apartheid South Africa, where the state placed many restrictions on Black educational opportunities. The restriction of oppressed groups to less skill–intensive jobs is consistent with my model, since less skill–intensive jobs are often complementary to more skill–intensive jobs.

Hurst et al. (2024) argue that managerial and customer-facing jobs are often reserved, while non-managerial and non-customer-facing jobs are often unreserved, due to tastes for discrimination

against members of oppressed groups who perform managerial or customer-facing tasks. The restriction of oppressed groups to non–management and non–customer–facing jobs is also consistent with my model, since management jobs are often complementary to non–management jobs and customer–facing jobs are often complementary to non–customer–facing jobs.

3.2 The optimal size of the set of reserved tasks

3.2.1 Preliminaries

Next I characterize the optimal size of the set of reserved tasks R. I begin with a preliminary result:

Proposition 2. *Suppose that assumption 1 holds. Then there exists a unique optimal value of R, which I denote by R^*.*

Proof. See appendix □

The proof of proposition 2 makes use of part 4 of assumption 1, which implies that w_d is strictly quasi-concave in R. Notice that quasi-concavity is a weaker condition than concavity and that w_d is not necessarily concave in R. However, strict quasi-concavity of w_d is sufficient to establish that w_d has a unique maximum in R.

3.2.2 The competition effect and the complementarity effect

In order to determine the optimal value of R, I decompose the effect of changing R on w_d into two components. Taking the derivative of (16) with respect to R yields:

$$\frac{\partial w_d}{\partial R} = \underbrace{\frac{\partial^2 F}{\partial L^2}\frac{\partial L}{\partial R}\left(L\frac{R}{\alpha_d}\right)^{1/\sigma}}_{competition\ effect} + \underbrace{\frac{1}{\sigma}\left(L\frac{R}{\alpha_d}\right)^{(1-\sigma)/\sigma}\left(\frac{\partial L}{\partial R}\frac{R}{\alpha_d}+\frac{L}{\alpha_d}\right)}_{complementarity\ effect} \tag{22}$$

Increasing R has two effects on w_d. First, increasing R reduces the effective labor supply of oppressed group workers, reducing competition from oppressed group workers for access to the non-labor factor of production Z. This effect, which I refer to as the competition effect, is the first term in (22). By assumption 1, $\partial^2 F/\partial L^2 < 0$, and it is straightforward to show that $\partial L/\partial R < 0$, so the competition effect is always positive. A stronger competition effect therefore implies a larger optimal value of R. Second, increasing R affects the degree to which dominant group workers benefit from complementary labor supplied by oppressed group workers. This effect, which I refer to as the complementarity

effect, is the second term in (22). The complementarity effect can be either positive or negative, and for sufficiently large R, the complementarity effect is always negative. A stronger complementarity effect therefore implies a smaller optimal value of R.

3.2.3 Labor abundance, labor scarcity, and discrimination

I apply the concepts of competition and complementarity effects to explain the effect of changing the quantity of the non–labor factor of production Z. Changes in the quantity of the non–labor factor affect the scarcity of labor relative to the non–labor factor. When Z is small, labor is abundant relative to the non-labor factor, while when Z is large, labor is scarce relative to the non–labor factor. Notice that changing Z is equivalent to changing the measures of both dominant and oppressed group workers α_d and α_o while keeping the ratio of dominant to oppressed group workers α_d/α_o fixed. Thus by studying the effect of changing Z, I study the effect of changing the scarcity of labor relative to the non–labor factor without changing the relative composition of the labor force. In order to study the effects of changing Z, I impose the following additional assumption:

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

1. $\partial^2 F/\partial Z \partial L > 0$
2. $\partial^3 F/\partial Z \partial L^2 > 0$

The first part of assumption 2 states that an increase in the quantity of the non-labor factor of production increases the marginal product of labor. The second part of assumption 2 states that an increase in the quantity of the non-labor factor of production reduces the curvature of the production function with respect to L. These condiations are satisfied by common production functions such as the CES prodution function.

Proposition 3. *Suppose that assumptions 1 and 2 hold, and suppose that the dominant group imposes discrimination. Then:*

1. *R^* is decreasing in Z, and strictly decreasing whenever $\alpha_d/(\alpha_d + \alpha_o) < R^* < 1$.*
2. *The wage ratio w_d^*/w_o^* is decreasing in Z, and strictly decreasing whenever $\alpha_d/(\alpha_d + \alpha_o) < R^* < 1$.*

Proof. See appendix. □

When labor is abundant, the competition effect is strong and so the dominant group optimally chooses to reserve a large number of tasks to reduce labor market competition from the oppressed group. When labor is scarce, the competition effect is weaker relative to the complementarity effect, and so the dominant group optimally chooses to reserve a smaller number of tasks. When fewer tasks are reserved, the wage gap between the dominant and oppressed groups decreases.

3.2.4 Discrimination or free labor markets?

If there are no costs to imposing discrimination, then the dominant group always prefers imposing discrimination to allowing free labor markets, because discrimination always increases the dominant group wage relative to free labor markets by proposition 1. However, when there is a fixed cost to imposing discrimination, then the dominant group may prefer to allow free labor markets rather than imposing discrimination. I now discuss how the fixed cost of discrimination affects the choice about whether to impose discrimination or to allow free labor markets.

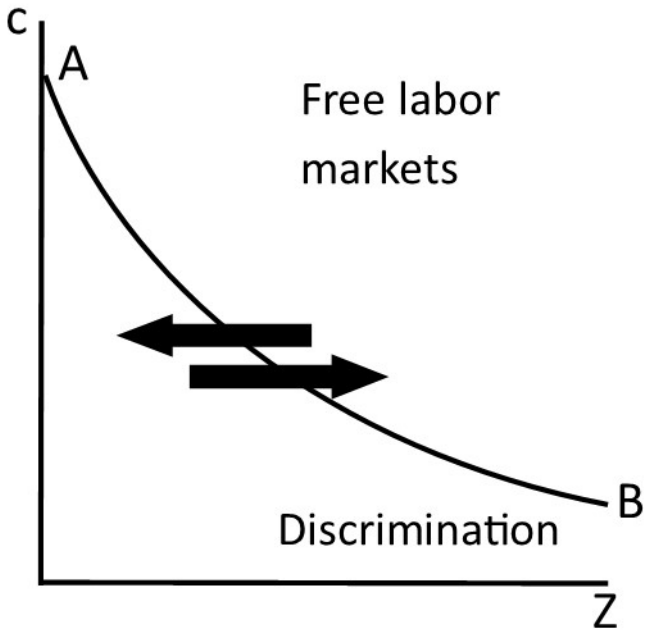
Proposition 4. *Suppose that assumptions 1 and 2 hold. Then given Z, there exists $\bar{c}(Z)$ such that the dominant group imposes discrimination if $c < \bar{c}(Z)$ and allows free markets if $c \geq \bar{c}(Z)$. Moreover, $\bar{c}(Z)$ is decreasing in Z.*

Proof. See appendix □

The dominant group imposes discrimination if the benefit of imposing discrimination in the form of higher wages is greater than the fixed cost of enforcing discrimination. The wage difference between discrimination and free labor markets decreases as Z increases, that is, as labor becomes more scarce, and so the critical level of c where the labor market regime switches is decreasing in Z.

Figure 4 presents a graphical summary of proposition 4. The downward-sloping curve AB is the critical value of c. When c is larger than the critical value, the dominant group prefers to allow free labor markets, while when c is below the critical value, the dominant group prefers to impose discrimination. Notice that a decrease in the scarcity of labor can cause a shift from free labor markets to discrimination, and conversely an increase in the scarcity of labor can cause a shift from discrimination to free labor markets. These shifts are represented by the arrows in figure 4. In section 4 below I argue that a decrease in the scarcity of labor led to the introduction of job reservations in apartheid South Africa in the 1920s and 1930s, and that a subsequent increase in the scarcity of labor led to the removal of job reservations in the 1970s and 1980s.

Figure 4: Optimal institutions



This figure shows how the optimal institutional regime for dominant group workers depends on the relative scarcity of labor, indexed by Z, and the cost of enforcing discrimination c. The arrows show how an increase in labor scarcity can cause a shift from discrimination to free labor markets, and conversely how a decrease in labor scarcity can cause a shift from free labor markets to discrimination.

3.3 What if assumption 1 does not hold?

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

Assumption 3. *For all* $L \in [\alpha_d, \alpha_d + \alpha_o]$,

$$\frac{\partial}{\partial L}\left(\frac{\partial F}{\partial L}L\right) = \frac{\partial^2 F}{\partial L^2}L + \frac{\partial F}{\partial L} < 0 \tag{23}$$

Assumption 3 states that for all feasible values of L, the total payment to labor is decreasing in aggregate labor L. Assumption 2 may hold if the elasticity of substitution between aggregate labor L and the non-labor factor of production Z is sufficiently small.

Using assumption 3, I can show the following:

Proposition 5. *Suppose that assumption 3 holds. Then* $R^* = 1$. *Any finite value of* σ *is optimal.*

Proof. See appendix. □

Proposition 5 shows that when the elasticity of substitution between labor and the non-labor factor of production is small (and hence assumption 3 is more likely to hold), the dominant group may choose to exclude the oppressed group from the labor market completely by setting $R = 1$. Intuitively, if the elasticity of substitution between labor and the non-labor factor of production is small, then competition for access to the non-labor factor of production is very harmful to the dominant group, and so the competition effect is strong. One way in which the dominant group may choose to exclude the oppressed group from the labor market completely is through ethnic cleansing or genocide.

It is likely that the elasticity of substitution between land and labor is lower than the elasticity of substitution between capital and labor. Thus, my model suggests that ethnic cleansing and genocide may be more likely to appear in societies in which the main non-labor factor of production is land, while institutionalized discrimination may be more likely to appear in societies in which the main non-labor factor of production is capital. For example, Esteban et al. (2015) argue that the Rwandan genocide was motivated by conflicts over access to land. In contrast, institutionalized discrimination in South Africa was largely a phenomenon of capital-intensive urban labor markets and the capital-intensive mining industry.

3.4 Effects of changing group sizes on wages under different labor market regimes

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

Proposition 6. *1. Under free labor markets, $dw/d\alpha_o < 0$.*

2. Suppose that assumption 1 holds. Then under optimal discrimination, $dw_d^*/d\alpha_o \geq 0$. If $R^* < 1$, then $dw_d^*/d\alpha_o > 0$.

Proof. See appendix. □

Proposition 6 states that if discrimination is optimal, then as the oppressed group becomes larger, the dominant group wage increases. Intuitively, if it is optimal to set $R^* < 1$, then the dominant group benefits on net from the complementary labor provided by the oppressed group, and so the

dominant group also benefits from increasing the size of the oppressed group. Proposition 6 implies that the dominant group may want 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. In section 3.2 below I show that not only did the apartheid government reject a policy of removing Blacks from the White economy through ethnic cleansing, but the apartheid government also instituted policies explicitly designed to increase Black participation in the White economy.

Like proposition 1, proposition 6 does not necessarily hold if the set of reserved tasks is exogenously fixed. Proposition 7 presents this result formally:

Proposition 7. *Suppose that σ and R are exogenously fixed. Then for any $R < 1$, there exists $\sigma < \infty$ such that $\partial w_d / \partial \alpha_o < 0$.*

Proof. See appendix. □

4 Applying the model to apartheid

My model suggests two empirically relevant predictions. The first prediction is that the choice between institutionalized discrimination and free labor markets depends on the relative abundance or scarcity of labor. When labor is abundant, discrimination is more likely, and when labor is scarce, free labor markets are more likely. The second prediction is that under discrimination, the dominant group benefits from increasing the labor supply of the oppressed group, and so the dominant group may enact policies designed to increase the labor supply of the oppressed group. In this section I discuss each of these predictions in the context of apartheid South Africa.

4.1 Labor abundance, labor scarcity, and the rise and fall of apartheid

Legally enforced job reservations were first introduced in South Africa in the 1920s, and 1930s, largely in response to a wave of migration of unskilled, mostly Afrikaans–speaking Whites from rural to urban areas within South Africa. These migrants were known at the time as "poor Whites".[5] Poor Whites competed for jobs with unskilled Blacks who had also migrated from rural to urban areas. In this context, a militant White labor movement developed, demanding state protection

[5]Nearly all histories of South Africa, including Hutt (1964), Lipton (1985), Seekings and Nattrass (2005), Feinstein (2005), and Thompson (2014), agree that the "poor White" problem was the fundamental cause of the introduction of job reservation.

for White labor against Black competition. The most dramatic manifestation of this militant labor movement was the Rand revolt of 1922, when plans by management to replace White workers with Black workers led to a strike that effectively turned into a rebellion against the state, and that had to be suppressed by 20,000 army troops using tanks, artillery, and aircraft. The strikers' slogan, "Workers of the world, unite and fight for a White South Africa," indicates the degree to which labor and racial politics were intertwined during this period (Feinstein (2005), p. 81).

In response to the White labor movement, throughout the 1920s and 1930s the South African government implemented legislation that imposed job reservations for Whites across progressively broader areas of the economy. Relevant legislation included the Industrial Concilation Act of 1924, which reserved many jobs for members of all-White unions, the Minimum Wages Act of 1925, which set high minimum wages in historically White occupations that effectively excluded Blacks, the Mines and Works Amendment Act of 1926, which reserved many mining industry jobs for Whites, and the Industrial Conciliation Act of 1937, which also used minimum wage rules to effectively exclude Blacks from many jobs. This process culminated with the victory of the National Party in the (racially segregated) election of 1948, which is usually considered to be the beginning of the apartheid era. The National Party represented relatively poor Afrikaans–speaking Whites, as opposed to richer English–speaking Whites who mostly supported the opposition United Party. Soon after gaining power, the National Party extended job reservations to all sectors of the economy.

Discrimination under apartheid was the most severe in the 1950s and 1960s. By the 1970s, the structure of the South African labor force had begun to change. In particular, by the 1970s improvements in White education ensured that nearly all White workers acquired at least some skills. In 1970 96.1% of Whites had attended school through at least standard 6 (equivalent to 8 years of education), compared to only 21.4% of urban Blacks and 6.6% of rural Blacks (Feinstein (2005), p. 161). However, increases in the demand for skill and the relatively small size of the White population ensured that while unskilled Black labor remained abundant, skilled labor was scarce.

Scarcity of skilled labor in the 1970s led to changes in the apartheid system. Many job reservations were relaxed during this period and Blacks were allowed to enter some skilled and semi-skilled jobs that had previously been reserved for Whites, a phenomenon known as the "floating color bar" (Mariotti, 2012).

In 1977 the South African government established the Wiehahn commission to study policies to improve the labor market. The Wiehahn commission identified scarcity of skilled labor as the core

problem of the South African economy. The commission's final report, issued in 1979, stated that as a result of the "ever-increasing process previous of industrialization... the already thinly stretched resources of skilled manpower in the country were placed under severe strain." The commission noted in particular that job reservations for Whites imposed restrictions "on the very category of workers [i.e. Blacks]... whose better training and utilisation are a *sine qua non* for the future economic growth and stability of the Republic" (Feinstein (2005), p. 241). The commission concluded by recommending the abolition of job reservations. The government followed the recommendations of the Wiehahn commission, removing most job reservations outside the mining industry by 1984 and most mining industry job reservations by 1988. It is noteworthy that the removal of legally enforced job reservations occurred in the context of continuing White political control, as South Africa's transition to full democracy and the enfranchisement of non-White voters did not occur until 1994.

Why were job reservations imposed and tightened in the 1920s, 1930s, and 1940s, and then relaxed and ultimately removed in the 1970s and 1980s? My model suggests that a key difference between the two periods was the relative scarcity of labor. In the 1920s and 1930s, the relative abundance of unskilled labor increased, in particular as a result of the migration of unskilled "poor Whites" to urban areas. In the context of labor abundance, protection of unskilled Whites from competition from unskilled Blacks through institutionalized discrimination was highly beneficial for Whites relative to a free labor market. As a result, institutionalized discrimination was imposed and the size of the set of reserved tasks was progressively increased. By the 1970s, nearly all White workers had at least some skills and so competition from unskilled Blacks was no longer relevant for Whites. Competition from skilled Blacks remained relevant to White workers, but in the 1970s and 1980s both White and Black skilled labor were scarce. In the context of labor scarcity, protection of skilled Whites from competition from skilled Blacks through institutionalized discrimination was less beneficial to Whites. As a result, the size of the set of reserved tasks was reduced, and ultimately discriminatory institutions were dismantled. Both of these effects are consistent with the predictions of my model.

4.2 Policies designed to increase Black labor supply under apartheid

As discussed in the previous subsection, the beginning of the apartheid era is usually dated to the victory of the National Party in 1948. 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.[6] The first program was known as "total apartheid".[7] 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.

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 program, and not the total apartheid program, was enacted into policy.[8] Proponents of baasskap accepted and supported the continuing growth of the Black population in South African cities and other White areas. The goal of proponents of baasskap was not to remove Blacks from the White economy but rather to increase inequality between Whites and Blacks by expanding and entrenching the job reservation policies that had been introduced in the 1920s and 1930s. 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

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

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

[8] 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.

policies associated with the baasskap faction were explicitly designed to increase Black labor supply to 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 (Gwaindepi and Siebrits, 2020; Feinstein, 2005).

Why did the apartheid state explicitly reject a policy of removing Blacks from the White economy, and why did the state impose policies explicitly designed to increase Black labor supply into the formal White economy, even as the state reserved many jobs for Whites? In my model, when discrimination is optimal, increasing the labor supply of oppressed group workers increases the wage of dominant group workers. Thus, dominant group workers may simultaneously support discrimination against oppressed group workers and policies designed to increase the labor supply of oppressed group workers.

5 Institutionalized discrimination against migrant workers in contemporary economies

In the contemporary world, the most important application of my model is to the political economy of migration, with native workers as the dominant group and migrant workers as the oppressed group. My model explains the common practice of admitting migrant workers, but only to work in certain jobs. In this section I apply my model to understanding policies related to migration in China, Japan, South Korea, and the United States.

5.1 China

Institutionalized discrimination in China is enforced through the hukou system, which is often described as a form of Chinese apartheid (e.g. The Economist (2014)). Under the hukou system, all Chinese workers are officially assigned to a place of residence. The most important distinction is between workers who are assigned to a rural hukou and workers who are assigned to an urban hukou. Children inherit their hukou status from their parents, and hukou status is only loosely related to the actual locations where workers live. In particular, many workers with rural hukou status migrate to urban areas to find work, with 80-100 million rural hukou holders working in urban areas in 1999

(Chan and Zhang, 1999).

In urban areas, workers with different hukou statuses perform different jobs. While the difference in occupations between workers with different hukou statuses is caused in part by differences in socio-economic characteristics between the two groups, it is also caused in part by legal barriers to employing people with rural hukou status in urban areas. Chan et al. (1999) (p. 428) explain the causes of occupational divergence between rural and urban hukou holders, writing, "The requirement to have a permit (based on local hukou status) to work in many urban jobs greatly limits the opportunities of non-hukou migrants. They are most likely to end up on the bottom rungs of the occupational hierarchy, and, typically, physically segregated from and socially marginalised by mainstream society."

Wang (2005) argues that the purpose of the hukou system is to benefit workers with urban hukou status at the expense of workers with rural hukou status in the context of an economy in which labor is abundant. He writes (p. 55), "This urban minority dominates China politically, economically, and culturally, and often uses its power to maintain and justify the PRC hukou system that gives it privilege and a sense of superiority." This argument is consistent with the idea in my model that job reservations (in this case for people with urban hukou status) serve to redistribute income towards members of politically powerful social groups at the expense of politically less powerful social groups.

5.2 Japan and South Korea

Consider two migration regimes: one in which migration is not allowed, and another in which migration is allowed but with migrant workers restricted to certain jobs. The first regime corresponds to discrimination with $R = 1$, that is with all jobs reserved for native workers, while the second regime corresponds to discrimination with $R < 1$, that is, with only some jobs reserved for native workers. Proposition 3, which shows that R is decreasing in the scarcity of labor, suggests that increasing scarcity of labor could cause a shift from the first regime to the second.

Japan and South Korea have both recently changed their immigration policies in ways that are consistent with this prediction. Historically both Japan and South Korea have had almost zero immigration. However, declining birth rates in both countries have led to labor scarcity, which has prompted both countries to implement guest worker programs. In each country, migrant workers are allowed to work only in certain jobs. The Japanese program, known as the Specified Skilled Worker program, allows migrants to work only in 16 specified occupational areas such as nursing,

construction, and agriculture. The Specified Skilled Worker program was introduced in 2019 and aims to admit 800,000 workers by 2029.[9] The South Korean program, known as the Employment Permit System, also restricts migrants to certain occupations also including construction and agriculture. The Employment Permit System was introduced in 2004 and admits around 56,000 workers per year.[10]

5.3 The United States

Many observers have noted that the US government does not seem to do as much as it 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. Immigration policy effectively restricts illegal immigrants to certain jobs such as domestic service and agricultural labor. In my model, citizens benefit from the presence of illegal immigrants when immigrants are subjected to optimal job restrictions, and so citizens are unlikely to support policies that reduce the size of the illegal immigrant population.

6 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 are assigned to endogenously created categories of "good" and "bad" jobs. 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 to maximize the wage of workers in the dominant group, and I characterize the optimal set of reserved tasks. I show that the both the choice between free labor markets and discrimination and the optimal severity of discrimination depend on the abundance or scarcity of labor. I also show that under optimal discrimination, increasing the size of the oppressed group benefits the dominant group.

[9] See https://www.mofa.go.jp/mofaj/ca/fna/ssw/us/overview/ for information on the Specified Skilled Worker program.

[10] See https://gsp.cgdev.org/legalpathway/employment-permit-system-eps/ for more information on the Employment Permit System.

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.

In the contemporary world, the most important application of my model is to understanding the politics of migration and the regulation of migrant labor in developed countries. Job reservations for native workers can build political support for migration among natives. Given the large potential benefits to migration, measures that increase political support for migration can have a large positive effect on global welfare. However, the similarity between job reservations for native workers and oppressive discriminatory systems such as apartheid raises difficult questions about the ethics, political philosophy, and law of job reservation. It seems likely that job reservations for native workers will become more prevalent in the future, and so an interdisciplinary research program studying both the positive and normative aspects of job reservations is called for.

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.

Acemoglu, D. and J. A. Robinson (2012). *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. New York: Crown Publishers.

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

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.

Chan, K. W., T. Liu, and Y. Yang (1999). Hukou and non-hukou migrations in china: Comparisons and contrasts. *International Journal of Population Geography 5*(6), 411–494.

Chan, K. W. and L. Zhang (1999). The hukou system and rural-urban migration in china: Processes and changes. *China Quarterly 160*, 818–855.

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

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.

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.

Harris, J. R. and M. P. Todaro (1970). Migration, unemployment and development: A two-sector analysis. *American Economic Review 60*(1), 126–142.

Hsieh, C.-T., E. Hurst, C. I. Jones, and P. J. Klenow (2019). The allocation of talent and u.s. economic growth. *Econometrica 87*(5), 1439–1474.

Hurst, E., Y. Rubinstein, and K. Shimizu (2024). Task-based discrimination. *American Economic Review 114*(6).

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

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.

Leon-Ledesma, M. A. and M. Satchi (2011). The choice of ces production techniques and balanced growth. *Working paper*.

Leon-Ledesma, M. A. and M. Satchi (2019). Appropriate technology and balanced growth. *Review of Economic Studies 86*(2), 807–835.

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

Lewis, W. A. (1954). Economic development with unlimited supplies of labour. *Manchester School 22*(2), 139–191.

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. (2012). Labor markets during apartheid in south africa. *Economic History Review 65*(3), 1100–1122.

McGee, D. (2024). Exploitation through racialization. *Working paper*.

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

Mukand, S. W. and D. Rodrik (2020). The political economy of liberal democracy. *Economic Journal 130*(627), 765–792.

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.

Norrell, R. J. (1986). Caste in steel: Jim crow careers in birmingham, alabama. *Journal of American History 73*(3).

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.

The Economist (2014). Ending apartheid. pp. April 19.

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

Wang, F.-L. (2005). *Organizing Through Division and Exclusion: China's Hukou System*. Stanford: Stanford University Press.

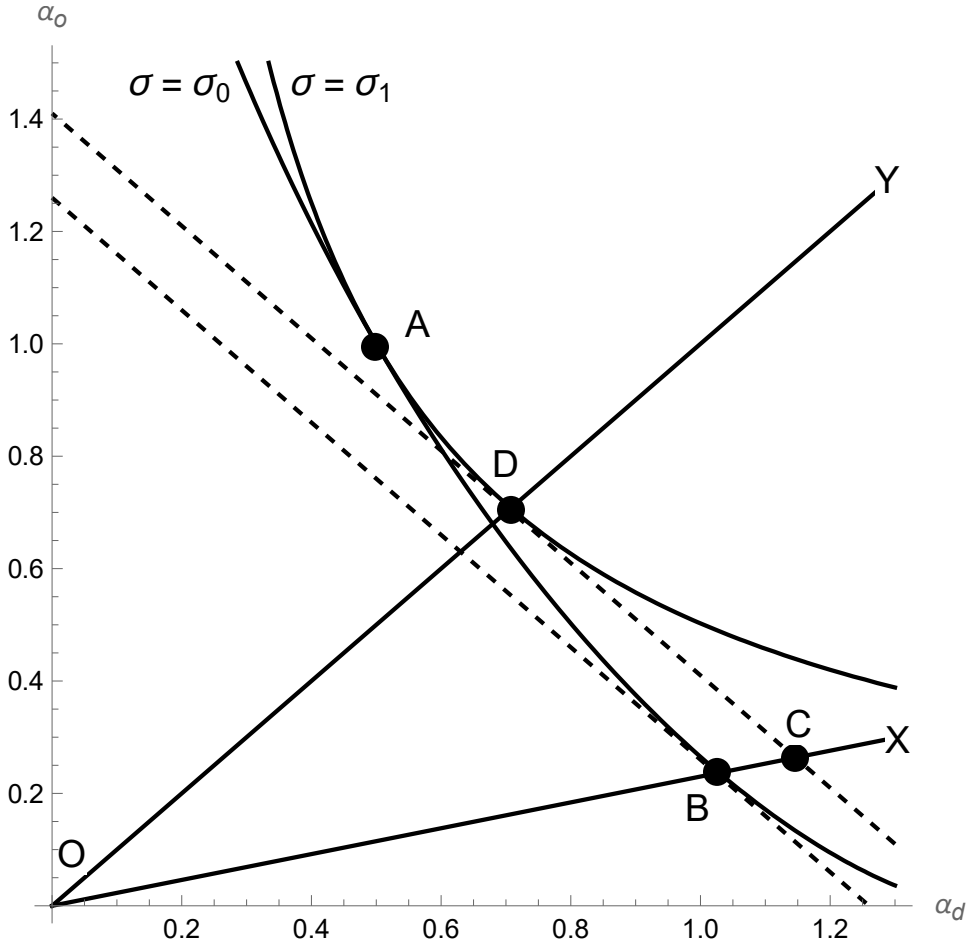
Wright, G. (1986). *Old South, New South: Revolutions in the Southern Economy since the Civil War*. Baton Rouge: Louisiana State University Press.

A Proofs

A.1 Proof of lemma 1

The proof of lemma 1 makes use of figure 5, 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 4. 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 $\sigma_0 > \sigma_1$.

Figure 5: Proof of Lemma 1



The ray OX is the set of points where $\alpha_d/(\alpha_d + \alpha_o) = R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0)$. 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 must be located below A, as depicted in figure 2. In addition, examination of (19) 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:

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

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

$$\hat{L}(\alpha_d^A, \alpha_o^A; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma_1) = \hat{L}(\alpha_d^D, \alpha_o^D; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma_1) \tag{25}$$

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,

$$\hat{L}(\alpha_d^B, \alpha_o^B; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) < \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) \tag{26}$$

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,

$$\hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) = \lim_{\sigma \to \infty} \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma) \tag{27}$$

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

$$\lim_{\sigma \to \infty} \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma) = \lim_{\sigma \to \infty} \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma) \tag{28}$$

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,

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

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,

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

Putting (24), (26), (27), (28), (29), (30), and (25) together in order yields:

$$\begin{aligned}
\hat{L}(\alpha_d^A, \alpha_o^A; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) &= \hat{L}(\alpha_d^B, \alpha_o^B; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) && (31)\\
&< \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) && (32)\\
&= \lim_{\sigma \to \infty} \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma) && (33)\\
&= \lim_{\sigma \to \infty} \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma) && (34)\\
&= \lim_{\sigma \to \infty} \hat{L}(\alpha_d^D, \alpha_o^D; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma) && (35)\\
&= \hat{L}(\alpha_d^D, \alpha_o^D; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma_1) && (36)\\
&= \hat{L}(\alpha_d^A, \alpha_o^A; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma_1) && (37)
\end{aligned}$$

This completes the proof of lemma 1.

A.2 Proof of proposition 2

By proposition 1, $\sigma = \underline{\sigma}$ at the optimum. Set $\sigma = \underline{\sigma}$ and rearrange (22) to get:

$$\frac{\partial w_d}{\partial R} = \frac{\partial L}{\partial R} \frac{R}{\alpha_d} \left(L \frac{R}{\alpha_d} \right)^{(1-\underline{\sigma})/\underline{\sigma}} \left\{ \frac{\partial^2 F}{\partial L^2} L + \frac{1}{\underline{\sigma}} \frac{\partial F}{\partial L} \left[1 + \frac{1}{\partial L / \partial R} \frac{L}{R} \right] \right\} \tag{38}$$

Use the expressions for $\xi(R)$ and $\zeta(R)$ from (14) and (15). Part 4 of assumption 1 states that either $\zeta(R) < 0$ for all R, or $\zeta(R)$ is strictly increasing in R. If $\zeta(R) < 0$ for all R then $\partial w_d / \partial R > 0$ for all R, so the unique maximum of w_d is $R^* = 1$. If $\zeta(R)$ is strictly increasing then there exists R^* such that $\zeta(R) < 0$ for $R < R^*$ and $\zeta(R) > 0$ for $R > R^*$, since $\lim_{R \to \alpha_d/(\alpha_d+\alpha_o)} \xi(R) = -\infty$, so $\zeta(R) < 0$ for sufficiently small R. Thus there exists R^* such that w_d is strictly increasing for all $R < R^*$ and strictly decreasing for all $R > R^*$. Therefore w_d is strictly quasi-concave in R, which implies that w_d has a unique maximum in R.

A.3 Proof of proposition 3

Consider again the expression for $\partial w_d/\partial R$, from (38), and define $\xi(R,Z)$ and $\zeta(R,Z)$ according to (14) and (15), explicitly noting the dependence of ξ and ζ on Z. Let $R^*(Z)$ be the optimal value of R for a given value of Z. Then $\zeta(R^*(Z),Z) = 0$ whenever $\alpha_d/(\alpha_d+\alpha_o) < R^* < 1$, and so $\xi(R^*(\hat{Z}),Z) > 0$ for all Z sufficiently close to $\hat{Z}$. By assumption 3, an increase in Z causes both $\partial F/\partial L$ and $\partial^2 F/\partial L^2$ to increase, and so an increase in Z causes $\zeta(R^*(\hat{Z}),Z$ to increase for all Z sufficiently close to $\hat{Z}$.

Consider two values of Z, $\bar{Z}$ and $\underline{Z}$, with $\bar{Z} > \underline{Z}$. From the above, $\zeta(R^*(\underline{Z}),\bar{Z}) > 0$ for $\bar{Z}$ sufficiently close to $\underline{Z}$. Since $\zeta(R^*(\bar{Z}),\bar{Z}) = 0$, and since $\zeta(R)$ is strictly increasing in R whenever $R^* < 1$ from the proof of proposition 2, it must be the case that $R^*(\bar{Z}) < R^*(\underline{Z})$. So R^* is strictly decreasing in Z whenever $\alpha_d/(\alpha_d+\alpha_o) < R^* < 1$.

From (19), an increase in R increases the wage ratio w_d/w_d. Thus the optimal wage ratio w_d^*/w_o^* is also strictly decreasing in Z whenever R^* is strictly decreasing in Z.

A.4 Proof of proposition 4

As in the previous proof, let $R^*(Z)$ be the optimal value of R for a given value of Z. The dominant group prefers to impose discrimination if:

$$(1-c)w_d(R^*(Z),Z) \geq w(\alpha_d+\alpha_o,Z) \tag{39}$$

Rewriting using (16) yields that the dominant group prefers to impose discrimination if:

$$\frac{\frac{\partial F(L(R^*(Z)),Z)}{\partial L}\left(L(R^*(Z))\frac{R^*(Z)}{\alpha_d}\right)^{1/\sigma}}{\frac{\partial F(\alpha_d+\alpha_o,Z)}{\partial L}} \geq \frac{1}{1-c} \tag{40}$$

Define

$$\Delta(R,Z) = \frac{\frac{\partial F(L(R),Z)}{\partial L}\left(L(R)\frac{R}{\alpha_d}\right)^{1/\sigma}}{\frac{\partial F(\alpha_d+\alpha_o,Z)}{\partial L}} \tag{41}$$

Choose $\underline{Z}$ and $\bar{Z}$ such that $\underline{Z} < \bar{Z}$. By assumption 2, $\partial^2 F/\partial L^2$ is increasing in Z. Since $L(R) < \alpha_d+\alpha_o$ this implies that:

$$\Delta(R^*(\bar{Z}),\underline{Z}) \geq \Delta(R^*(\bar{Z},\bar{Z})) \tag{42}$$

In addition, since $R^*(\underline{Z}$ is the optimal level of R when $Z = \underline{Z}$,

$$\Delta(R^*(\underline{Z}), \underline{Z}) > \Delta(R^*(\bar{Z}), \underline{Z})) \tag{43}$$

Therefore $\Delta(R^*(Z), Z)$ is strictly decreasing in Z, which in turn implies that results stated in the proposition.

A.5 Proof of proposition 5

From (19), the wage ratio w_d/w_o is increasing in R. Differentiating L with respect to R shows 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.

A.6 Proof of proposition 6

By assumption 1, $dw/d\alpha_o < 0$.

Differentiate (16) with respect to α_o and apply the envelope theorem to get:

$$\frac{dw_d^*}{d\alpha_o} = \frac{\partial w_d}{\partial \alpha_o} = \frac{\partial L}{\partial \alpha_o}\frac{R}{\alpha_d}\left(L\frac{R}{\alpha_d}\right)^{(1-\sigma)/\sigma}\left[\frac{\partial^2 F}{\partial L}L + \frac{1}{\sigma}\frac{\partial F}{\partial L}\right] \tag{44}$$

Recall again the expression for $\xi(R)$ derived in (14). Since $\partial L/\partial R < 0$ for all $\sigma < \infty$, $\xi(R) < 1$. Using this fact, comparing (44) with (38) shows that whenever (38) is equal to zero, (44) is strictly greater than 0. If $R^* < 1$, then (38) is equal to zero at $R = R^*$. Therefore, if $R^* < 1$, $dw_d^*/d\alpha_o > 0$.

If $R^* = 1$ then it is straightforward to verify that $dw_d^*/d\alpha_o = 0$.

A.7 Proof of proposition 7

The sign of (44) is the same as the sign of the expression within the square brackets in (44). By assumption 1, for σ sufficiently large, the expression in the square brackets is negative. Therefore $\partial w_d/\partial \alpha_o < 0$ for sufficiently large σ.