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

James P. Choy

May 24, 2024

Abstract

Institutions in some societies force employers to discriminate. I develop a theory of institutionalized discrimination. Optimal discrimination sorts workers from different social groups into complementary tasks. Workers in the politically dominant social group benefit from complementary labor supplied by oppressed group workers, but are harmed by competition from oppressed group workers for access to non-labor factors of production. The tradeoff between these two forces determines whether ethnic cleansing, institutionalized discrimination, or free labor markets are optimal for workers in the dominant group. 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, inflicted either by spontaneously formed mobs or by more organized groups such as the Ku Klux Klan. 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 two social groups, a dominant group and an oppressed group. Workers from each group can choose to work in any of a number of different tasks. Final output is produced from an aggregate of labor applied to different tasks and from a non-labor factor of production. The dominant group may use its political power to impose labor market regulations reserving some subset of the available tasks for members of the dominant group. 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, perhaps because the median voter in the dominant group is a worker (rather than an owner of the non-labor factor).

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. Under discrimination, oppressed group workers then affect dominant group workers in two ways. First, oppressed group workers benefit dominant group workers by providing complementary labor, increasing the productivity of dominant group workers. Second, oppressed group workers harm dominant group workers by competing with dominant group workers for access to the non-labor factor of production. The tradeoff between these two forces determines the optimal size of the set of reserved tasks. Krueger (1963) and Porter (1978) only discuss one side of the tradeoff, the harm to dominant group workers from competition with oppressed group workers for access to the non-labor factor of production. Mariotti (2012) only discusses the other side of the tradeoff, the benefit to dominant group workers from complementary labor provided by oppressed group workers.[1] Lundahl (1982) implicitly includes both sides of the tradeoff, but Lundahl (1982) does not make the tradeoff explicit and does not derive any of the consequences of the tradeoff discussed below.

I use the tradeoff between the costs and benefits of oppressed group workers to dominant group workers to discuss the determinants of three possible institutional regimes, namely ethnic cleansing and genocide, institutionalized discrimination, and free labor markets. More specifically, I present the following results. First, I show that in some cases, it is optimal for the dominant group to reserve all tasks, effectively excluding members of the oppressed group from the labor market completely. These cases correspond to ethnic cleansing or genocide. In other cases, it is optimal for the dominant group to leave some tasks unreserved. These cases correspond to institutionalized discrimination in societies such as apartheid South Africa and the Jim Crow South, in which Black workers were a crucial part of the labor force even though Black workers were constrained relative to White workers. Second, I show that when not all tasks are reserved, the dominant group wage is increasing in the

[1] Another difference between my paper and Mariotti (2012) is that Mariotti (2012) studies the effect of discrimination on relative wages between the dominant and oppressed groups, while my paper studies the effect of discrimination on the absolute dominant group wage. It seems more natural that dominant group workers would seek to maximize absolute rather than relative wages, but studying absolute wages is much more difficult from a technical perspective. One of the main contributions of this paper is to provide a general framework for studying the effect of task discrimination on absolute dominant group wages.

size of the oppressed group. This result implies that the dominant group may be willing to expend resources to increase the size of the oppressed group, for example by promoting immigration (or preventing emigration) by members of the oppressed group. Third, I show that the size of the set of reserved tasks and the wage gap between groups depend on the abundance or scarcity of labor relative to the non-labor factor of production. When labor is abundant, the size of the set of reserved tasks and the wage gap between groups are larger than when labor is scarce. I interpret this result as showing that discrimination is more severe when labor is abundant. The difference between the dominant group wage under optimal discrimination and the free market wage is also higher when labor is abundant. If there is a fixed cost of imposing discriminatory institutions, then the dominant group is more likely to impose discrimination when labor is abundant, and the dominant group is more likely to allow a free labor market when labor is scarce.

I apply my model to the context of apartheid South Africa. At various points in the history of White political control of South Africa, all three of ethnic cleansing of Blacks, institutionalized discrimination against Blacks, and free labor markets for Whites and Blacks were either proposed or actually implemented as policies. I argue that my model helps to explain these policy choices. In addition I argue that my model can help to explain the racial division of labor across firms and the return to capital under apartheid.

I conclude by presenting examples of institutionalized discrimination in societies other than South Africa. I discuss the US South under Jim Crow, contemporary China, Saudi Arabia, Malaysia, Singapore, Japan, and South Korea. I argue that all of these sets of institutions have features broadly consistent with my model. Many of the examples in this section relate to migration, with migrant workers forming the oppressed group and native workers forming the dominant group. In these cases my model helps to explain the political economy of migration and regulations on migrant labor. In particular, my model shows that under optimal job reservation for native workers, increasing the number of migrant workers benefits native workers. This is true even in cases in which unrestricted migration would harm native workers. Thus, job reservations for native workers can build political support for immigration among natives. As birthrates in developed countries decline, labor migration from developing countries to developed countries is likely to become more important. At the same time, increasing political backlash against migration suggests that large-scale unrestricted migration is likely to be politically infeasible. Therefore I 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.

My model suggests a number of factors that determine which tasks are likely to be reserved, including among other things the prediction that reserved and unreserved tasks are likely to be complements. Other authors have suggested different factors that may explain which tasks are reserved. Mariotti (2012) argues that skill-intensive tasks are likely to be reserved, while less skill-intensive tasks are likely to be unreserved. Hurst et al. (2022) argue that management and customer-facing tasks are likely to be reserved, while non-management, non-customer-facing tasks are likely to be unreserved. My model is compatible with these other theories. Skill-intensive tasks are often complementary to less skill-intensive tasks, so the observation that skill-intensive tasks are often reserved while less skill-intensive tasks are often unreserved is consistent with my model. Similarly, management tasks are often complementary to non-management tasks, and customer-facing tasks are often complementary to non-customer-facing tasks, so the observation that management tasks and customer-facing tasks are often reserved while non-management and non-customer-facing tasks are often unreserved is also consistent with my model. Relative to Mariotti (2012) and Hurst et al. (2022), a contribution of my model is that my model makes sharp predictions not only about the elasticity of substitution between reserved and unreserved tasks but also about the size of the set of reserved tasks. These predictions have empirical counterparts that are not predicted by Mariotti (2012) or Hurst et al. (2022).

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) applies 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 Theory

2.1 The production technology

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

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

j. The output $q(j)$ of division j is produced according to the CES function:

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

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

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

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

Here τ_2 is the elasticity of substitution between divisions.

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

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

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

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

Instead of allowing a free market, the dominant group can reserve some subset of tasks for dominant group workers. Suppose that the dominant group wants to reserve a set of tasks with

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

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

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

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

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

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

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

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

The output of the unreserved divisions is:

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

Aggregate labor supply is:

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

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

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

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

2.1.1 An example

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 under Jim Crow, 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

As in the previous subsection, society consists of two groups, a dominant group and an oppressed group. The measures of the dominant and oppressed groups are α_d and α_o, respectively. Final output is a function of aggregate labor supply L and some other factor of production Z:

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

There are two possible labor market regimes, a free labor market and institutionalized discrimination. In the free labor market, following the example in the previous subsection, $L = \alpha_d + \alpha_o$. The wage for both groups in the free market is then

$$w = \frac{\partial F}{\partial L} \tag{10}$$

Under institutionalized discrimination, following the example in the previous subsection, 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 supply function L, with a discrete set of possible values of σ. For the remainder of the paper, I abstract from the specific technology suggested 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 supply is sufficiently large. For example, if the elasticity of substitution between the non-labor factor and labor is greater than 1, then the labor share of output is increasing in L. By assumption total output also increases in L, so the total payment to labor must be increasing in L. Even if the elasticity of substitution between Z and L is below 1, part 3 of assumption 1 may be satisfied if the elasticity of total output with respect to aggregate labor supply 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.

The wage of workers in the dominant group is $w_d = \partial F/\partial \alpha_d = (\partial F/\partial L)(\partial L/\partial \alpha_d)$, and the wage of workers in the oppressed group is $w_o = \partial F/\partial \alpha_o = (\partial F/\partial L)(\partial L/\partial \alpha_o)$. The following expression for the dominant group wage is useful:

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

Notice that the dominant group wage under discrimination is equal to the marginal productivity of aggregate labor $\partial F/\partial L$ multiplied by an additional factor $(LR/\alpha_d)^{1/\sigma} > 1$. This additional factor represents the degree to which the productivity of dominant group workers is increased by complementary labor supplied by oppressed group workers under discrimination.

The objective of the dominant group is to maximize the dominant group wage. To achieve this objective, the dominant group first chooses whether to allow a free labor market or to impose institutionalized discrimination. If the dominant group imposes institutionalized discrimination, then the dominant group chooses R and σ to solve:

$$\max_{R,\sigma} w_d(\alpha_d, \alpha_o, R, \sigma) \tag{17}$$

subject to the constraint (12).

One reason why the dominant group may choose to maximize the dominant group wage, as opposed to total dominant group income or some weighted average of labor and non-labor income for members of the dominant group, is if the median voter in the dominant group is a worker rather than a capital or land owner. In apartheid South Africa, the standard view is that the balance of political power was held by poor White workers, who imposed discrimination to benefit themselves at the expense of Black workers and possibly also at the expense of 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.

2.3 Results

2.3.1 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 supply function, first proposed by de La Grandville (1989). The marginal rate of technical substitution (MRTS) between α_d and α_o is

$$MRTS = \frac{\partial L/\partial \alpha_d}{\partial L/\partial \alpha_o} = \left[\frac{R}{(1-R)}\frac{\alpha_o}{\alpha_d}\right]^{1/\sigma} \tag{18}$$

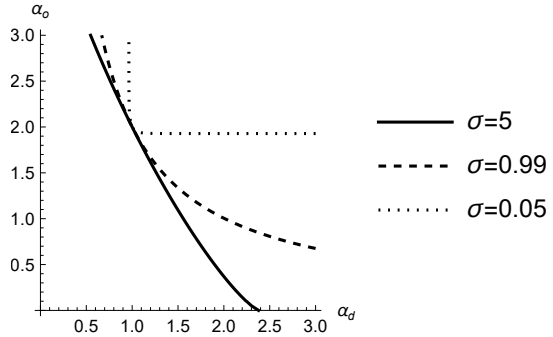
The MRTS is also equal to the wage ratio w_d/w_o. Let $\bar{\alpha}_d$ and $\bar{\alpha}_o$ be particular values of α_d and α_o and let $\bar{\mu}$ a particular MRTS. Then I can define a family of normalized CES aggregate labor supply that all have MRTS $\bar{\mu}$ at the point $(\bar{\alpha}_d, \bar{\alpha}_o)$,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

$$\left.\frac{\partial L/\partial \alpha_d}{\partial L/\partial \alpha_o}\right|_{\bar{\alpha}_d, \bar{\alpha}_o, R(\bar{\alpha}_d, \bar{\alpha}_o, \bar{\mu}, \sigma)} = \bar{\mu} \tag{19}$$

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

$$\hat{L}(\alpha_d, \alpha_o; R(\bar{\alpha}_d, \bar{\alpha}_o, \bar{\mu}, \sigma), \sigma) = \left[(R(\bar{\alpha}_d, \bar{\alpha_o}, \bar{\mu}, \sigma))^{1/\sigma}\alpha_d^{(\sigma-1)/\sigma} + (1 - R(\bar{\alpha}_d, \bar{\alpha}_o, \bar{\mu}, \sigma))^{1/\sigma}\alpha_o^{(\sigma-1)/\sigma}\right]^{\sigma/(\sigma-1)} \tag{20}$$

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



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

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

I use these definitions to prove the following lemma:

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

Proof. See appendix. □

Lemma 1 states that when R is varied to hold the wage ratio w_d/w_o fixed, aggregate labor supply 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.*

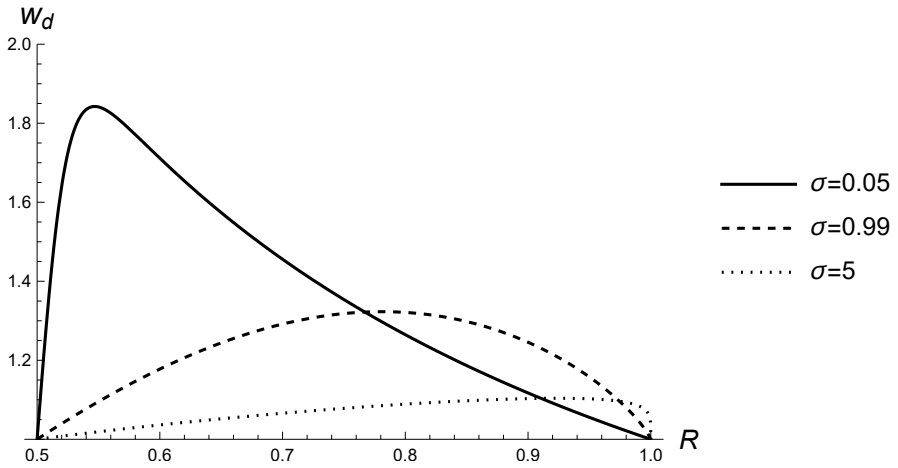
Proof. From lemma 1, decreasing σ while varying R to hold the wage ratio w_d/w_o fixed increases aggregate labor supply L. Assumption 1 states that increasing aggregate labor supply L increases the total payment to labor. If the total payment to labor increases while the wage ratio remains fixed, the wage of the dominant group must increase. Therefore, it is always possible to increase the dominant group wage by reducing 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. □

The first part of proposition 1 states that under institutionalized discrimination members of different social groups are assigned to tasks that are as complementary as possible. This result formalizes the idea that discrimination forcibly sorts members of different social groups into tasks that are as different as possible. The second part of proposition 1 states that under optimal discrimination the dominant group wage w_d is greater than the free market wage, so dominant group workers benefit from imposing discrimination relative to allowing a free labor 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 2 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 2: 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 $\sigma = 5$

2.3.2 Discussion

Proposition 1 states that under discriminatory institutions, the elasticity of substitution between reserved and unreserved tasks is as low as possible. This theory does not fully pin down the set of reserved tasks, since there may be many ways to partition the set of tasks that generate similar elasticities of substitution between reserved and unreserved tasks. When there are multiple ways to partition the set of tasks that are consistent with my theory, other factors may also affect the partition of tasks. One additional factor that may affect the partition of tasks is the difference in skill intensity between tasks. Highly skill-intensive tasks are often complementary to less skill-intensive tasks. For example, highly skilled doctors are complementary to less skilled nurses. Thus my theory suggests that one occupation out of the set {doctors, nurses} is likely to be reserved, but

my theory makes no prediction about which occupation is likely to be reserved. My theory does not make a prediction in this case because my theory abstracts from skill differences between groups. In practice, members of the dominant social group often have higher skill levels than members of the oppressed group. When members of different social groups have different skill levels, it is efficient to assign members of the more skilled social group to the more skill-intensive task and members of the less skilled social group to the less skill-intensive task. Thus, skill differences between groups can help to explain why in practice high-skilled occupations such as doctors are more likely to be reserved than lower-skilled occupations such as nurses.

Another factor that may affect the partition of tasks is tastes for discrimination linked to specific tasks, as suggested by Hurst et al. (2022). Hurst et al. (2022) argue that members of the dominant group have a taste for discrimination against workers from the oppressed group who perform managerial or customer-facing tasks. Managers are complementary to non-managerial workers, so my theory predicts that one occupation out of the set {managers, non-managers} is likely to be reserved, but again my theory does not predict which occupation is likely to be reserved. Tastes for discrimination (as well as differences in skill content) may help to explain why in practice managerial occupations seem to be reserved more often than non-managerial occupations. Similarly, many customer-facing tasks are complementary to non-customer facing tasks. For example, waiters and dishwashers are complements in the production of restaurant meals. Thus, my theory predicts that one occupation out of the set {waiters, dishwashers} is likely to be reserved, but not which occupation is likely to be reserved. Tastes for discrimination may help to explain why the tasks of waiters seem to be more likely to be reserved than the tasks of dishwashers.

2.3.3 The optimal size of the set of reserved tasks

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

Proposition 2. *Suppose that assumption 1 holds. Then:*

1. *There exists a unique optimal value of R, which I denote by R^*.*
2. *For all $0 < \underline{\sigma} < \infty$, $R^* > \alpha_d/(\alpha_d + \alpha_o)$.*
3. *Let w_d^* and w_o^* be the dominant and oppressed group wages given the optimal values of R and σ. Then for all $\underline{\sigma} < \infty$, $w_d^*/w_o^* > 1$.*

Proof. See appendix □

The proof of the first part 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. The second part of proposition 2 states that for all $\underline{\sigma} < \infty$, the ratio of reserved to unreserved tasks is greater than the ratio of dominant group members to oppressed group members, and the dominant group wage is greater than the oppressed group wage. Thus, under optimal discrimination, the oppressed group is constrained relative to the free market.

Next I decompose the effect of changing R on w_d into two components. Taking the derivative of (16) with respect to R yields:

$$\frac{\partial w_d}{\partial R} = \underbrace{\frac{\partial^2 F}{\partial L^2}\frac{\partial L}{\partial R}\left(L\frac{R}{\alpha_d}\right)^{1/\sigma}}_{competition\ effect} + \underbrace{\frac{1}{\sigma}\left(L\frac{R}{\alpha_d}\right)^{(1-\sigma)/\sigma}\left(\frac{\partial L}{\partial R}\frac{R}{\alpha_d}+\frac{L}{\alpha_d}\right)}_{complementarity\ effect} \tag{21}$$

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 (21). By assumption 1, $\partial^2 F/\partial L^2 < 0$, and $\partial L/\partial R < 0$, 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 (21). 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.

I use the competition and complementarity effects to characterize the optimal size of the set of reserved tasks R^*. It is particularly interesting to consider the conditions under which $R^* = 1$. When $R^* = 1$, it is optimal for the dominant group to exclude the oppressed group from the labor market completely. One of the most effective ways to exclude a group from the labor market completely is by physically removing the group from society, for example through ethnic cleansing or genocide. Thus, the case where $R^* = 1$ can be interpreted as the case in which ethnic cleansing or genocide are

optimal for the dominant group. Proposition 3 states some of the conditions under which $R^* = 1$:

Proposition 3. *1. If $F(Z, L) = L$, then $R^* < 1$ for all $\underline{\sigma} < \infty$.*

2. Suppose that F satisfies assumption 1. Then there exists $\bar{\sigma} > 1$ such that for all $\underline{\sigma} < \bar{\sigma}$, $R^* < 1$.

3. Suppose that F satisfies assumption 1. Then for $\underline{\sigma}$ sufficiently large, $R^* = 1$.

Proof. See appendix. □

The first part of proposition 3 discusses the case in which the non-labor factor of production is irrelevant, in which case the competition effect does not matter. The proposition states that in this case it is never optimal to set $R = 1$. Thus, competition for access to the non-labor factor of production is a necessary condition for ethnic cleansing or genocide to be optimal. The idea that genocide is motivated by competition over access to non-labor factors of production is the essence of the theory of genocide presented by Esteban et al. (2015). The second and third parts of proposition 4 state that it is more likely that $R = 1$ is optimal when $\underline{\sigma}$ is large. When $\underline{\sigma}$ is large, dominant group and oppressed group labor are not very complementary and so the complementarity effect is weak relative to the competition effect, increasing the optimal level of R.

A special case of the third part of proposition 3 is the limit as $\underline{\sigma}$ approaches infinity, in which case the complementarity effect becomes non-existent. In this limit, it is effectively impossible to assign members of different groups to economically distinct tasks and so the only feasible options are a free labor market in which members of the dominant and oppressed groups participate on equal terms or ethnic cleansing that completely eliminates the oppressed group. The third part of proposition 3 shows that in this case ethnic cleansing is always optimal. Conversely, when it is feasible to assign members of different social groups to distinct tasks, the dominant group may prefer not to ethnically cleanse the oppressed group, even when ethnic cleansing is feasible. In section 3.2 below I show that in apartheid South Africa, a policy of ethnically cleansing Blacks was explicitly proposed, but rejected, in favor of a policy that allowed Blacks to continue to work in the White economy in segregated tasks.

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

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

$$\frac{\partial}{\partial L}\left(\frac{\partial F}{\partial L}L\right) = \frac{\partial^2 F}{\partial L^2}L + \frac{\partial F}{\partial L} < 0 \tag{22}$$

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

Using assumption 2, I can show the following:

Proposition 4. *Suppose that assumption 2 holds. Then $R^* = 1$. Any finite value of σ is optimal.*

Proof. From (18), 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. □

Comparing propositions 3 and 4 shows that excluding the oppressed group from the labor market completely by setting $R = 1$ is more likely when the elasticity of substitution between labor and the non-labor factor of production is small (and hence assumption 2 is more likely to hold), while allowing the oppressed group to retain some access to the labor market is more likely when the elasticity of substitution between labor and the non-labor factor of production is large (and hence assumption 1 is more likely to hold). 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. In this case it may be optimal to set $R = 1$. On the other hand, if the elasticity of substitution between labor and the non-labor factor of production is large, then competition from the oppressed group for access to the non-labor factor of production is relatively unimportant, and so the competition effect is weak relative to the complementarity effect. In this case it is more likely to be optimal to set $R < 1$.

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.

2.3.4 Effects of changing group sizes on wages under discrimination

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

Proposition 5. *Suppose that assumption 1 holds. Then* $dw_d^*/d\alpha_o \geq 0$. *If* $R^* < 1$, *then* $dw_d^*/d\alpha_o > 0$.

Proof. See appendix. □

Proposition 5 states that if $R^* < 1$, 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, despite competition from the oppressed group for access to the non-labor factor of production. If the oppressed group provides net benefits to the dominant group, then the dominant group also benefits from increasing the size of the oppressed group. Proposition 5 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 5 does not necessarily hold if the set of reserved tasks is exogenously fixed. Proposition 6 presents this result formally:

Proposition 6. *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. □

The proof of proposition 6 shows that for any $R < 1$, $dw_d/d\alpha_o < 0$ when σ is sufficiently large. Intuitively, when σ is very large, institutionalized discrimination is approximately equivalent to a free labor market. In a free labor market, the only effect of increasing the size of the oppressed group is to increase competition for access to the non-labor factor of production, so in a free labor market increasing the size of the oppressed group reduces the wage of the dominant group.

2.3.5 Labor abundance, labor scarcity, and discrimination

Next I consider effects of changing the quantity of the non-labor factor of production Z. Changes in the quantity of the non-labor factor of production affect the relative scarcity of labor. When Z is small, labor is relatively abundant, while when Z is large, labor is relatively scarce. In order to study the effects of changing Z, I impose the following additional assumption:

Assumption 3. *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 3 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 3 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. This condition is satisfied by common production functions such as the CES prodution function.

Proposition 7. *Suppose that assumptions 1 and 3 hold, and suppose that* $\underline{\sigma} < \infty$ *and* $R^* < 1$. *Then:*

1. R^* *is strictly decreasing in* Z.
2. *The wage ratio* w_d^*/w_o^* *is strictly decreasing in* Z.

Proof. See appendix. □

Proposition 7 states that when labor is scarce relative to the non-labor factor of production, optimal discrimination is less severe. Intuitively, when labor is scarce, there is not much competition for access to the non-labor factor of production. In this case the competition effect is weak relative to the complementarity effect, and so it is optimal to set R relatively small. In contrast, when labor is abundant, competition for access to the non-labor factor of production is very costly to the dominant group. In this case the competition effect is strong relative to the complementarity effect, and so it is optimal to set R relatively large. Larger values of R increase the wage ratio between the dominant group and the oppressed group.

Next I consider the case in which there is a fixed cost to imposing and maintaining discriminatory institutions. I formalize this case as assumption 4:

Assumption 4. *In order to impose institutionalized discrimination, the dominant group must pay a fixed cost f. The cost is divided equally among all dominant group workers, so the amount paid by each worker in the dominant group is f/α_d. Let $t = f/\alpha_d$ be the per-capita cost of imposing discrimination. The dominant group maximizes the consumption of dominant group workers, which is equal to the free market wage w in a free labor market and equal to $w_d - t$ under institutionalized discrimination.*

Proposition 8. *Suppose that assumptions 1, 3, and 4 hold. Then either the dominant group chooses to allow a free labor market for all Z, or the dominant group imposes discrimination for all Z, or there exists $\bar{Z}$ such that for $Z < \bar{Z}$ the dominant group imposes discrimination, while for $Z \geq \bar{Z}$ the dominant group allows a free labor market.*

Proof. See appendix □

Relative to the free market, imposing institutionalized discrimination benefits the dominant group by assigning members of the oppressed group to complementary tasks and also by reducing competition from the oppressed group for access to the non-labor factor of production. The first benefit is independent of Z, but the second benefit is smaller when Z is large. Thus the benefit of discrimination relative to the free market for members of the dominant group is decreasing in Z. If there is a fixed cost to imposing discrimination, the dominant group is thus less likely to impose discrimination when Z is large.

3 Applying the model to apartheid

In this section I apply the results of my model to understanding the institutions of apartheid South Africa.

3.1 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 the problem of "poor Whites".[2] The poor White population consisted primarily of Afrikaans-speaking migrants from rural to urban areas who were largely unskilled and in many cases

[2]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.

almost completely lacked formal education. Poor Whites competed for jobs with unskilled Blacks who had also migrated from rural to urban areas. In 1932, the Carnegie Commission estimated that about one-third of Afrikaans-speaking Whites were "poor Whites"; in turn, Afrikaans-speakers were about two-thirds of the total White population (Feinstein (2005), p. 85). In this context, a militant White labor movement developed, demanding state protection 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.

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 skilled labor was scarce. During the 1970s many job reservations were relaxed 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, both White and Black unskilled labor were abundant, as indicated by the low wages and poverty experienced by unskilled "poor Whites". In this context 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 while unskilled Black labor remained abundant, 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 skilled labor was scarce. In this context protection of skilled Whites from competition from skilled Blacks through institutionalized discrimination was less beneficial. 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 the model discussed in section 2.3.5

3.2 Ethnic cleansing versus discrimination under early 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.[3] The first program was known as "total apartheid".[4] Proponents of total apartheid proposed to expel Blacks from White areas of South Africa, including South Africa's cities, the best agricultural areas, and the areas containing the largest mineral deposits, and to split the territory of South Africa into separate, independent, racially homogeneous states for Blacks and Whites. Had it been implemented, this program would have completely removed Black workers from the White economy. The total apartheid program corresponds to the idea in my model that in some cases, the dominant group can benefit by completely excluding the oppressed group from the labor market, since completely excluding the oppressed group from the labor market minimizes competition from the oppressed group for access to non-labor factors of production.

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.[5] 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

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

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

[5]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.

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 policies associated with the baasskap faction were explicitly designed to increase the amount of Black participation in the formal White economy. For example, South African tax and land use policy was explicitly designed to force Blacks to seek formal employment in the White economy by forcing Blacks to acquire currency and by making traditional forms of herding and subsistence agriculture infeasible (Gwaindepi and Siebrits, 2020; Feinstein, 2005). Regarding tax policy, Van Der Horst (1942) (p. 111) writes, "Taxes were levied for different reasons. Some were imposed... with the definite object of forcing Natives to work for Europeans." Similarly, Van Der Horst (1942) describes various legislative measures prohibiting Blacks from renting land from Whites. The purpose of this legislation was explicitly to force Blacks to provide wage labor on commercial farms as opposed to engaging in subsistence agriculture; Van Der Horst (1942) (p.291) explains the purpose of the legislation as "a means of assisting farmers to obtain labour."

The rejection of the total apartheid program of ethnic cleansing and the adoption of the baasskap program allowing continued integration of Black labor into the White economy is consistent with the argument in section 2.3.3 that when it is feasible to assign dominant and oppressed group workers to economically distinct tasks, it can be optimal for the dominant group to allow the oppressed group to participate in the labor market by setting $R < 1$. Baasskap policies designed to increase Black participation in the White economy are consistent with the argument in section 2.3.4 that when $R < 1$, increasing the size of the oppressed group benefits dominant group workers.

3.3 Socially heterogeneous firms

Section 2.1 argues that in the free market, the allocation of workers from different social groups to tasks is indeterminate. In particular, if there are multiple firms, it is consistent with the free market

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

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

response to the perception that underground mining jobs would soon be monopolized by Blacks. As Feinstein (2005), p. 88, puts it, "It was clear [to the Mining Regulations Commission] that Africans must be prevented from performing such work, not because they lacked the competence to do it but, on the contrary, precisely because they were, or soon would be, competent. A new colour bar act was urgently required."

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

3.4 The return to capital under discrimination

Consider the following proposition:

Proposition 9. *In the limit as $\underline{\sigma}$ approaches 0,*

1. *The size of the set of reserved tasks R approaches $\alpha_d/(\alpha_d + \alpha_o)$.*
2. *Aggregate labor supply approaches $L = \alpha_d + \alpha_o$, which is the same as aggregate labor supply in the free market.*
3. *The return to the non-labor factor of production Z approaches the return in the free market.*

Proof. See appendix. □

Proposition 9 shows that for $\underline{\sigma}$ approaching 0, discrimination does not cause labor misallocation and hence does not reduce the return to the non-labor factor of production. Of course, in reality it is unlikely that $\underline{\sigma} = 0$, and if $\underline{\sigma} >$) then discrimination does reduce the return to the non-labor factor. However, proposition 9 can be interpreted as showing that the distortionary effects of discrimination may be relatively small. Intuitively, discrimination has two effects on the return to the non-labor factor. First, by increasing the wage for reserved tasks, discrimination reduces the return to the non-labor factor. Second, by reducing the wage for unreserved tasks, discrimination increases the

return to the non-labor factor. The second effect partially offsets the first effect, reducing the overall negative effect of discrimination on the return to the non-labor factor.

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

4 Institutionalized discrimination in other societies

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

4.1 The US South under Jim Crow

Wright (1986) discusses job reservation in the US South under Jim Crow. He finds that when Black and White workers performed similar jobs, they received similar wages. However, there were many jobs that were effectively barred to Blacks. He writes (p. 185), "Job discrimination in the better-paying positions was far more important than wage differentials for the same job. Blacks could get the going wage in the unskilled market, but there was a virtual upper limit to their possible progress above that level." Job discrimination was particularly notable in positions which required

skills learned on the job. White workers could expect to achieve these positions after several years of experience, but Black workers could not. Wright writes (p. 185), "In Birmingham, for example, with one of the largest concentrations of black industrial labor, more than 80 percent of black workers who stayed for ten years made no upward progress at all toward better jobs during this time. By contrast, half of white workers moved up after ten years." The fact that even experienced Black workers could not advance suggests that Black exclusion from higher paying jobs was not due to lack of skill.

Wright points out that unlike in apartheid South Africa, job discrimination in the US South was for the most part not enforced by law. He suggests that statistical discrimination may explain the observed disparities between Blacks and Whites. I argue, however, that at least part of the observed job discrimination in the South was enforced by informal institutions through the threat of racial violence. Consider, for example, the practice of "whitecapping", which was prevalent throughout the South in the early part of the Jim Crow era. In the 1972 Mississippi statute banning the practice, whitecapping is defined as "threats, direct or implied, of injury to the person or property of another, to intimidate such a person into an abandonment or change of home or employment." Whitecapping threats were often directed at Black workers who took jobs that had customarily been held by Whites. An example of whitecapping comes from the Supreme Court case Hodges v. United States, decided in 1906. The case concerned a group of Whites who had threatened violence to force Black workmen to leave their jobs in a lumber mill in Poinsett County, Arkansas. This kind of violence enacted a form of job reservation, since Poinsett County had a significant Black population who were able to work in other jobs free from violence. The court ruled that the federal government did not have jurisdiction to enforce laws against whitecapping, effectively legalizing whitecapping throughout the South, since all-White Southern state juries were very unlikely to convict Whites of crimes against Blacks. Whitecapping declined in the later part of the Jim Crow era. However, other forms of racial violence developed to threaten Blacks who took jobs that had customarily been reserved for Whites, and employers who offered those jobs. For example, in 1943 the Alabama Dry Dock and Shipping Company (ADDSCO) promoted twelve Black workers to the position of welder, a job that previously had been held exclusively by Whites, although there were nearly 7,000 Black workers in other jobs within ADDSCO. The next day 4,000 Whites rioted throughout the shipyard, closing the shipyard for a week, causing 50 injuries, and requiring army troops to quell the violence (Nelson, 1993).

4.2 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." For urban hukou holders, he writes (p. 27), "The PRC hukou system has contributed very significantly to the impressive economic growth and development of technological sophistication in China, a typical dual-economy nation featuring massive unskilled and surplus labor."

4.3 Saudi Arabia

The private sector economy in Saudi Arabia is dominated by non-citizens, mainly consisting of migrants from other Arab countries and from South and South-East Asia. In 2011 non-citizens made

up 90% of the private non-oil sector workforce (Peck, 2017). Prior to 2011, the Saudi state made some efforts to encourage citizen employment in the private sector, but these efforts were largely ineffective. This changed in 2011 when the state initiated the Nitaqat program, which imposed quotas limiting the number of non-citizens that firms could hire, combined with harsh penalties if the quotas were not met. Nitaqat effectively forces Saudi firms to discriminate in favor of citizens and against non-citizens. Nitaqat quotas vary by industry, so that tasks associated with low-quota industries are more likely to be effectively reserved for citizens than tasks associated with high-quota industries. Peck (2017) shows that Nitaqat substantially increased private-sector Saudi employment, at the cost of increasing the exit rates of affected firms and reducing total employment in those firms. Notably, Nitaqat did not reduce the total number of migrant workers in Saudi Arabia, although it did shift the distribution of migrant workers across firms (Thiollet, 2022).

The introduction of the Nitaqat program in Saudi Arabia was motivated by the Arab Spring pro-democracy protests beginning in late 2010 (Thiollet, 2022). It seems that the Arab Spring protests increased the political bargaining power of poor Saudi citizens, forcing the Saudi government to implement the Nitaqat program to benefit poor Saudi citizens, possibly at the expense of Saudi elites. This is similar to the way in which the introduction of discrimination in South Africa was motivated by the Rand revolt of 1922 and the associated labor movement, as discussed in section 3.1 above.

4.4 Malaysia and Singapore

Migrant workers make up 15% of the labor force in Malaysia and 38% of the labor force in Singapore.[6] The majority of these migrants are unskilled and semiskilled workers from lower income countries. In Malaysia, the largest sources of migrant workers are Indonesia, Bangladesh, and Nepal. In Singapore, a large number of migrant workers also come from Malaysia and China. In both countries, migrant workers are admitted on temporary work visas tied to a specific employer, and migrant workers are not allowed to change employers without government permission. There are restrictions on employment of migrant workers that vary by sector, and in some cases migrants are restricted to specific jobs. For example, in Singapore, a maximum of 35% of the employees of a firm in the services sector can be migrant workers, but up to 83% of the employees of a firm in the construction sector

[6]For statistics on migrant workers in Singapore, see https://www.mom.gov.sg/foreign-workforce-numbers. For Malaysia, see World Bank (2020).

can be migrant workers.[7] In Malaysia, migrant workers in the services sector are permitted to work only as cooks, in cleaning and sanitation, on island resorts, in spas, as reflexologists, in hotels in back office positions only, as golf caddies, or in cargo handling.[8]

In 2020, in response to the Covid pandemic, the Malaysian government tightened restrictions on the types of jobs that migrant workers could hold by banning migrant workers from employment in the manufacturing and services sectors. Migrant workers could continue to work only in construction or agriculture. The stated motive for this change was to benefit citizen workers by increasing citizen employment.[9] The observation that a negative shock to labor demand due to the Covid pandemic led to an increase in the size of the set of jobs reserved for Malaysian citizens is consistent with my model.[10]

4.5 Discussion

Many of the examples in this section relate to migration, with migrant workers forming the oppressed group and native workers forming the dominant group. In this context, my model shows that under optimal job reservation for natives, increasing the number of migrants benefits native workers. In contrast, when the labor market is free, increasing the number of migrants harms native workers. Thus, job reservations for natives can build political support among natives for migration.

As developed country populations age and shrink relative to populations in less-developed countries, migration from less-developed countries to developed countries is likely to grow in importance. At the same time, increasing political backlash against unrestricted migration suggests that large-scale unrestricted migration is likely to be politically infeasible. Thus I predict that in the future, job reservations for native workers will become more prevalent. Already, a number of countries have begun to experiment with large-scale migration combined with job reservations for natives of the type already established in countries like Saudi Arabia, Malaysia, and Singapore. For example, Japan and South Korea have historically had almost zero immigration. However, both countries have

[7] See https://www.mom.gov.sg/passes-and-permits/work-permit-for-foreign-worker/ for rules about migrant workers in Singapore.

[8] See https://www.moha.gov.my/index.php/en/bahagian-pa-dasar-dasar-semasa/sektor-yang-dibenarkan for rules about migrant workers in Malaysia.

[9] See https://www.malaymail.com/news/malaysia/2020/07/29/putrajaya-limits-foreign-workers-only-to-construction-agriculture-plantatio/1889109 for details. Accessed April 26, 2024.

[10] Malaysian law also discriminates between ethnically Malay citizens and ethnically Chinese and Indian citizens, as discussed for example by Fang and Norman (2006). One aspect of this discriminatory regime is that ethnic Malay citizens are strongly preferred over ethnic Chinese and ethnic Indian citizens for public sector employment. This form of discrimination could also be analyzed as an example of my theory, although the distinction between public and private sector employment raises additional complications that are not fully addressed in my model.

recently created programs for migrant workers in which migrants are restricted to working in certain jobs. The relevant programs are the Employment Permit System in South Korea and the Specified Skilled Worker program in Japan, which admit migrants to work only in specific occupations such as agriculture, nursing care, and construction. The Korean program was introduced in 2004 and admits approximately 56,000 workers per year, while the Japanese program was introduced in 2019 and was expanded in 2024 to allow for over 800,000 workers in each five-year period. Details of these and other similar programs can be found at https://gsp.cgdev.org/.

5 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 work in different 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 use my model to explain the determinants of three possible institutional regimes, namely ethnic cleansing and genocide, institutionalized discrimination, and free labor markets.

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 work-

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

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

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

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

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.

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

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

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

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

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

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

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

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 (2022). Task-based discrimination. *working paper*.

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

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

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

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

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

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

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

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

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.

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

Van Der Horst, S. T. (1942). *Native Labour in South Africa*. Oxford: Oxford University Press.

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

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

World Bank (2020). Who is keeping score? estimating the number of foreign workers in malaysia. *Policy report*.

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

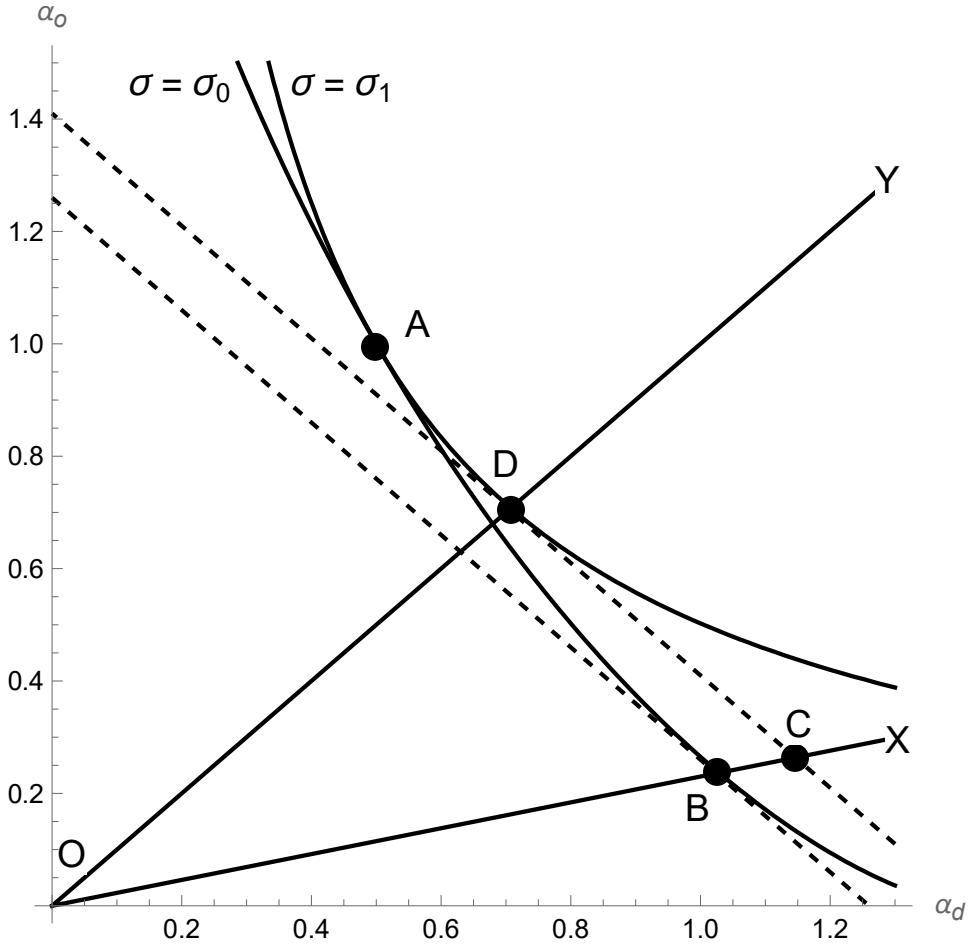
A Proofs

A.1 Proof of lemma 1

The proof of lemma 1 makes use of figure 3, 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 3. Fix a value of the MRTS $\bar{\mu}$, with $\bar{\mu} > 1$. The figure shows the isoquants of two members of the family of CES aggregate labor supply functions that have slope $\bar{\mu}$ at point A, with elasticities of substitution σ_0 and σ_1, and $\sigma_0 > \sigma_1$.

The ray OX is the set of points where $\alpha_d/(\alpha_d + \alpha_o) = R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0)$. 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

Figure 3: Proof of Lemma 1



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 (18) 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:

$$\hat{L}(\alpha_d^A, \alpha_o^A; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) = \hat{L}(\alpha_d^B, \alpha_o^B; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) \tag{23}$$

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

$$\hat{L}(\alpha_d^A, \alpha_o^A; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma_1) = \hat{L}(\alpha_d^D, \alpha_o^D; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma_1) \tag{24}$$

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,

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

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,

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

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,

$$\lim_{\sigma \to \infty} \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma) = \lim_{\sigma \to \infty} \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma) \tag{27}$$

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,

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

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,

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

Putting (23), (25), (26), (27), (28), (29), and (24) together in order yields:

$$\hat{L}(\alpha_d^A, \alpha_o^A; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) = \hat{L}(\alpha_d^B, \alpha_o^B; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) \tag{30}$$

$$< \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma_0) \tag{31}$$

$$= \lim_{\sigma \to \infty} \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_0), \sigma) \tag{32}$$

$$= \lim_{\sigma \to \infty} \hat{L}(\alpha_d^C, \alpha_o^C; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma) \tag{33}$$

$$= \lim_{\sigma \to \infty} \hat{L}(\alpha_d^D, \alpha_o^D; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma) \tag{34}$$

$$= \hat{L}(\alpha_d^D, \alpha_o^D; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma_1) \tag{35}$$

$$= \hat{L}(\alpha_d^A, \alpha_o^A; R(\alpha_d^A, \alpha_o^A, \bar{\mu}, \sigma_1), \sigma_1) \tag{36}$$

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 (21) to get:

$$\frac{\partial w_d}{\partial R} = \frac{\partial L}{\partial R} \frac{R}{\alpha_d} \left(L \frac{R}{\alpha_d} \right)^{(1-\underline{\sigma})/\underline{\sigma}} \left\{ \frac{\partial^2 F}{\partial L^2} L + \frac{1}{\underline{\sigma}} \frac{\partial F}{\partial L} \left[1 + \frac{1}{\partial L / \partial R} \frac{L}{R} \right] \right\} \tag{37}$$

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.

If $\underline{\sigma} > 0$ and $R \leq \alpha_d/(\alpha_d+\alpha_o)$, then the wage under discrimination w_d is equal to the free market wage. If $\underline{\sigma} < \infty$, then by proposition 1 the dominant group wage under optimal discrimination is greater than the free market wage. Therefore if $0 < \underline{\sigma} < \infty$, then $R \leq \alpha_d/(\alpha_d + \alpha_o)$ is not optimal.

A.3 Proof of proposition 3

I begin by deriving an expression for $\partial L/\partial R$ by differentiating (11) with respect to R:

$$\frac{\partial L}{\partial R} = \frac{1}{\sigma - 1}\left[R^{1/\sigma}\alpha_d^{(\sigma-1)/\sigma} + (1-R)^{1/\sigma}\alpha_o^{(\sigma-1)/\sigma}\right]^{1/(\sigma-1)}\left[\left(\frac{\alpha_d}{R}\right)^{(\sigma-1)/\sigma} - \left(\frac{\alpha_o}{1-R}\right)^{(\sigma-1)/\sigma}\right] \quad (38)$$

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

$$\left.\frac{\partial L}{\partial R}\right|_{R=1} = \frac{1}{\underline{\sigma} - 1}\alpha_d \quad (39)$$

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

$$\lim_{R\to 1}\frac{\partial L}{\partial R} = -\infty \quad (40)$$

Now consider the expressions for $\partial w_d/\partial R$ and $\zeta(R)$ derived in (37) and (15). Since the sign of w_d is the opposite of the sign of $\zeta(R)$ and since w_d is strictly quasi-concave in R by the proof of proposition 2, $R = 1$ is optimal if and only if $\lim_{R\to 1}\zeta(R) < 0$. If $\underline{\sigma} < 1$ then plugging (39) into (15) yields

$$\lim_{R\to 1}\zeta(R) = \frac{\partial^2 F}{\partial L^2}L + \frac{\partial F}{\partial L} \quad (41)$$

In this case, by assumption 1, $\lim_{R\to 1}\zeta(R) > 0$ and so $R = 1$ is not optimal.

If $\sigma > 1$, then taking the limit of (15) and using (40) yields:

$$\lim_{R\to 1}\zeta(R) = \frac{\partial^2 F}{\partial L^2}L + \frac{1}{\underline{\sigma}}\frac{\partial F}{\partial L} \quad (42)$$

By assumption 1, there exists $\bar{\sigma} > 1$ such that for $\underline{\sigma} < \bar{\sigma}$, $\lim_{R\to 1}\zeta(R) > 0$, and so $R = 1$ is not optimal. On the other hand, again by assumption 1, for $\underline{\sigma}$ sufficiently large, $\lim_{R\to 1}\zeta(R) < 0$, and so $R = 1$ is optimal.

A.4 Proof of proposition 5

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

$$\frac{dw_d^*}{d\alpha_o} = \frac{\partial w_d}{\partial \alpha_o} = \frac{\partial L}{\partial \alpha_o}\frac{R}{\alpha_d}\left(L\frac{R}{\alpha_d}\right)^{(1-\sigma)/\sigma}\left[\frac{\partial^2 F}{\partial L}L + \frac{1}{\sigma}\frac{\partial F}{\partial L}\right] \quad (43)$$

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 (43) with (37) shows that whenever (37) is equal to zero, (43) is strictly greater than 0. If $R^* < 1$, then (37) 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.5 Proof of proposition 6

The sign of (43) is the same as the sign of the expression within the square brackets in (43). 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.

A.6 Proof of proposition 7

Consider again the expressions for $\partial w_d/\partial R$, from (37), and define $\xi(R, Z)$ and $\zeta(R, Z)$ according to (14) and (15), explicitly noting the dependence of ξ and ζ on Z. Consider two values of Z, $\bar{Z}$ and $\underline{Z}$, with $\bar{Z} > \underline{Z}$. Let $R^*(Z)$ be the optimal value of R for a given value of Z. Then $\zeta(R^*(Z), Z) = 0$ whenever $R^* < 1$. By assumption 3, an increase in Z causes both $\partial F/\partial L$ and $\partial^2 F/\partial L^2$ to increase. Therefore, $\zeta(R^*(\underline{Z}), \bar{Z}) > 0$. 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.

From (18), 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.

A.7 Proof of proposition 8

Define

$$s = \frac{\alpha_d w_d}{\alpha_d w_d + \alpha_o w_o} \tag{44}$$

Then s is the share of total labor income that is paid to the dominant group. Notice that s can be written as

$$s = \frac{1}{1 + \frac{w_o}{w_d}\frac{\alpha_o}{\alpha_d}} \tag{45}$$

By (18), w_o/w_d is a function of R, so I can write $s = s(R)$.

Let $L(R)$ be aggregate labor supply when the size of the set of reserved tasks is R and the

elasticity of substitution between reserved and unreserved tasks is $\underline{\sigma}$. Define

$$\pi(R, Z) = \frac{\partial F(Z, L(R))}{\partial L} L(R) \tag{46}$$

Then $\pi(R, Z)$ is the total payment to labor. The sum of wages paid to the dominant group is then $s(Z)\pi(R, Z)$.

Notice that the sum of wages paid to the dominant group is the free market is:

$$\frac{\alpha_d}{\alpha_d + \alpha_o} \pi(\frac{\alpha_d}{\alpha_d + \alpha_o}, Z) \tag{47}$$

Define

$$\Delta(R, Z) = s(R)\pi(R, Z) - \frac{\alpha_d}{\alpha_d + \alpha_o} \pi(\frac{\alpha_d}{\alpha_d + \alpha_o}, Z) \tag{48}$$

Then $\Delta(R, Z)$ is the amount by which the sum of wages paid to the dominant group increases by moving from a free labor market to discrimination.

Finally, define $R^*(Z)$ to be the optimal value of R for a given Z. Consider two values of Z, $\underline{Z}$ and $\bar{Z}$, with $\underline{Z} < \bar{Z}$. Consider the following expressions:

$$\Delta(R^*(\bar{Z}), \bar{Z}) = s(R^*(\bar{Z})\pi(R^*(\bar{Z}), \bar{Z}) - \frac{\alpha_d}{\alpha_d + \alpha_o} \pi(\frac{\alpha_d}{\alpha_d + \alpha_o}, \bar{Z}) \tag{49}$$

$$\Delta(R^*(\bar{Z}), \underline{Z}) = s(R^*(\bar{Z})\pi(R^*(\bar{Z}), \underline{Z}) - \frac{\alpha_d}{\alpha_d + \alpha_o} \pi(\frac{\alpha_d}{\alpha_d + \alpha_o}, \underline{Z}) \tag{50}$$

Since $\partial F / \partial L$ is increasing in Z, we have

$$\frac{\alpha_d}{\alpha_d + \alpha_o} \pi(\frac{\alpha_d}{\alpha_d + \alpha_o}, \bar{Z}) - \pi(R^*(\bar{Z}), \bar{Z}) > \frac{\alpha_d}{\alpha_d + \alpha_o} \pi(\frac{\alpha_d}{\alpha_d + \alpha_o}, \underline{Z}) - \pi(R^*(\bar{Z}), \underline{Z}) \tag{51}$$

Thus,

$$\Delta(R^*(\bar{Z}), \underline{Z}) > \Delta(R^*(\bar{Z}), \bar{Z}) \tag{52}$$

Moreover, since $R^*(\bar{Z})$ is not optimal when $Z = \underline{Z}$,

$$\Delta(R^*(\underline{Z}), \underline{Z}) > \Delta(R^*(\bar{Z}), \underline{Z}) \tag{53}$$

Therefore $\Delta(R^*(Z), Z)$ is decreasing in Z, which implies the conclusions in the text of the

proposition.

A.8 Proof of proposition 9

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