Will Affirmative-Action Policies Eliminate Negative Stereotypes?

By STEPHEN COATE AND GLENN C. LOURY*

A key question concerning affirmative action is whether the labor-market gains it brings to minorities can continue without it becoming a permanent fixture in the labor market. We argue that this depends on how the policy affects employers' beliefs about the productivity of minority workers. We study the joint determination of employer beliefs and worker productivity in a model of statistical discrimination in job assignments. We prove that, even when identifiable groups are equally endowed ex ante, affirmative action can bring about a situation in which employers (correctly) perceive the groups to be unequally productive, ex post. (*JEL* D63, D82, J71)

> "I have a dream that my four little children will one day live in a nation where they will not be judged by the color of their skin but by the content of their character."
>
> —Martin Luther King Jr.
> (Washington, DC, August 1963)

Affirmative action is an important and controversial policy used to combat differences between groups in earnings and employment. Its pros and cons have been studied by scholars in many fields. Within economics the major focus of research has been on determining the importance of affirmative action for explaining improvements in the black–white earnings ratio since the 1960's (see e.g., Jonathan S. Leonard, 1984; James P. Smith and Finis Welch, 1984; Welch, 1989). An equally significant question, hitherto ignored by economists, is whether labor-market gains due to affirmative action can be expected to continue without it becoming a permanent fixture in the labor market.

An important component of this question would seem to be the impact of affirmative action on employers' stereotypes about the capabilities of minority workers. If affirmative action serves to break down negative stereotypes, then to the extent that these underlie discrimination, a temporary program of affirmative action should lead to permanent gains for minorities.[1] But if negative views about a minority group are not eroded or, indeed, are worsened by affirmative action, then it must be maintained permanently for that group's gains to be protected. Popular discussions of affirmative action often focus on just this issue. Advocates say that preferential policies break down negative views about minority workers by allowing them to demonstrate their capa-

*Coate: Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104-6297; Loury: Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215. Helpful comments on this work have been made by the referees, as well as by colleagues at Boston, Cornell, Harvard, Northwestern, and Stanford Universities, and the Universities of Chicago, Michigan, and Pennsylvania. The financial support of the University of Pennsylvania Research Foundation and the Bradley Foundation is gratefully acknowledged.

[1]We stress that the issue of stereotypes is but one aspect of the question of whether affirmative action must be a permanent measure. Though we will focus on negative stereotypes as the basis for discrimination, there are other kinds of discrimination. Some employers may simply refuse to promote workers from a certain group, even though they view them as equally capable. Then ongoing regulation of hiring patterns would be needed until employers' tastes are changed by greater exposure to this group in high-level positions or by the pressures of competition. Alternatively, if discrimination stems from cultural differences, affirmative action may encourage employers to try to assimilate these differences.

bilities. Critics say that affirmative action forces employers to lower standards, with the consequence that subsequent poor performance by preferred workers will only reinforce negative prejudices.

This paper offers a framework for the analysis of this issue by studying how the introduction of affirmative-action policy impacts on employers' beliefs about the capabilities of minority workers. We propose a job-assignment model in which employers observe the group identity, but not the productivity, of their workers. Employers form beliefs about the correlation between group identity and productivity which, in the equilibria of our model, must be correct. If workers in one group are seen as less productive, we say that employers have *negative stereotypes* about that group. We examine whether a policy of affirmative action can be expected to dispel these stereotypes. We thus shed light on the question of when such a policy is consistent with the eventual attainment of a color-blind society.

In our model, an employer who harbors negative stereotypes against some group is less likely to assign workers belonging to that group to the more highly rewarded jobs within the firm. This lowers the expected return for these workers on investments which make them more productive in such jobs. For this reason it is possible that employers' negative beliefs about a group are confirmed in equilibrium, even when all groups are *ex ante* identical. In this sense, negative stereotypes constitute a "self-fulfilling prophecy." This framework is a natural one for thinking about the problem because it allows employers' beliefs to be determined by their experience, while making that experience the result of the endogenous choices of workers. If affirmative action is to have any chance of changing employers' negative beliefs, these beliefs must be responsive to new evidence. Moreover, it is also necessary that minority workers respond to the enhanced opportunities created by affirmative action by producing evidence of greater productivity.

With this theory of stereotypes in hand, we consider the effects of affirmative action. We model this as a government-mandated constraint on employers requiring them to assign workers from each group to more rewarding jobs at the same rate. We ask whether the introduction of such a constraint is sufficient to induce employers, in the resulting equilibrium, to believe that workers' productivities are uncorrelated with their group identity.

Our results are mixed, providing credence to the views of those on both sides of the issue. There do exist circumstances under which affirmative action will necessarily eliminate negative stereotypes. However, there are also equally plausible circumstances under which minority workers continue to be (correctly) perceived as less capable, despite the affirmative-action constraint. Indeed, the policy can actually worsen employers' perceptions of the productivity of initially disadvantaged workers. This result is particularly striking, given our maintained hypothesis that the groups of workers are *ex ante* identical.

The reason that affirmative action may sometimes fail is simple. If employers continue to hold onto negative views about a group of workers then, to comply with the affirmative-action mandate, they must lower the standard used for assigning these workers to the better jobs within the firm. Lowering the standard may reduce investment incentives, however, because the favored workers see themselves as likely to succeed without acquiring the relevant skills. Thus, employers' negative stereotypes can continue to be confirmed in the equilibrium under affirmative action if they *patronize* the disadvantaged group—that is, if, believing a group to be less productive, they respond to the equal-representation constraint by making it easier for the less skilled workers in this group to succeed.

We also show that this logic has more general implications. First, it implies that if groups are unequally endowed *ex ante*, with employers having a realistic but not negatively stereotypic view of workers' productivity in the less endowed group, then the use of affirmative action may cause the *ex post* gap in group performance to widen. Second, it suggests that a policy which rewards workers directly for their economic advancement, rather than encouraging or forcing employers to promote them, will be a

more reliable way to eliminate negative stereotypes.

This paper is related to a large literature on employment discrimination. The two main theories of discrimination are a theory based on tastes, pioneered by Gary S. Becker (1957), and a statistical theory, studied initially by Kenneth J. Arrow (1973) and Edmund S. Phelps (1972).[2] Our paper builds on the statistical literature, being close in spirit to Arrow's work. Statistical models rely on imperfect observability of an employee's productivity to account for employers' use of group identity in their assessments. While Phelps assumed available measures of productivity to be noisier for minority workers,[3] Arrow showed that statistical discrimination can occur even when there are no such unexplained group differences. He noted that when employee productivity is endogenous, employers' prejudicial beliefs can be self-fulfilling.

In Arrow's (1973) model, employers offer lower wages to minorities for the same work, in equilibrium. We modify his setup so that workers receive equal pay for equal work, but minorities may have a lower probability of being assigned to the higher-paying jobs. This provides a theory of discrimination in job assignment rather than wages, unlike most previous work in this field.[4] Discriminatory wages for the same work is a flagrant violation of equal-employment laws, and relatively easy to detect. Discrimination in job assignment, which affirmative action seeks to counteract, is a more subtle phenomenon. It seems appropriate to ground an analysis of affirmative-action policy on the assumption that employers might discriminate in job assignment but not in wages.

Surprisingly, not much theoretical work has been done on affirmative action. Two papers which should be noted are Welch (1976) and Lundberg (1991).[5] Welch (1976) studies employment quotas in a model where discrimination is taste-based. He focuses on the economy-wide impact of affirmative action, particularly when the policy applies to some sectors but not others. He shows that affirmative action may result in unskilled workers being assigned to skilled jobs, or vice versa. Lundberg (1991) considers the problem of enforcing equal-opportunity laws when regulators do not observe firms' personnel policies and are uncertain about the link between workers' characteristics and their productivity. She notes some interesting differences in the effects of two regulatory regimes, one requiring wages to depend on a given set of worker characteristics in the same way for each group, and the other specifying that wages cannot be based on variables which may serve as proxies for race and sex. Neither of these papers takes up our main concern here—whether affirmative-action policy must be maintained permanently to assure the persistence of minority gains.

[2]While these are the main views, they are not the only ones. Michael A. Spence (1974) discusses a theory based on "signaling," where the relationship between education and ability is perceived to be different for different groups. Drawing on the sociolinguistic literature, Kevin Lang (1986) also offers a "language" theory of discrimination.

[3]Dennis J. Aigner and Glen G. Cain (1977), George J. Borjas and Matthew S. Goldberg (1979), Shelly J. Lundberg and Richard Startz (1983), and Lang (1990) have also presented models in this vein.

[4]One exception is Paul Milgrom and Sharon Oster (1987). They develop a model of discrimination in job assignments based on the idea that a firm may prevent the market from learning what it knows about the abilities of some workers by "hiding" them in less visible, lower-paying jobs. Coate and Sharon Tennyson (1992) present a model of discrimination in job assignment in their analysis of the impact of labor-market discrimination on self-employment.

[5]Loury (1987) presents a theoretical argument intended to justify the use of affirmative action, in the context of a model in which an individual's earnings ability is influenced by the community where he grows up. Racial segregation among communities may result in long-run differences in the distribution of outcomes between groups, even when both groups are equally able. Affirmative action represents one way of tackling these differences. Lawrence M. Kahn (1991) shows that affirmative-action policies have different effects in general-equilibrium models of taste discrimination depending on the source of prejudice (i.e., customer, employer, or co-worker). Andrew Schotter and Keith Weigelt (1992) present an interesting experimental study of the effect of bias in tournaments on effort levels. Milgrom and Oster (1987) also explore the impact of employment quotas in their model of discrimination in job assignment.

The next section introduces our model, and explains how negative, self-confirming stereotypes arise in equilibrium. Section II introduces the affirmative-action constraint on employers' behavior and establishes a sufficient condition for it to eliminate stereotypes. It is also shown in an illustrative example that, if this condition is not satisfied, affirmative action need not eliminate negative stereotypes and may in fact make them worse. Section III states and proves the main theorem, showing that our negative result does not depend on the special features of the example. We offer further policy discussion in Section IV and conclude in Section V.

I. Self-Fulfilling Negative Stereotypes

We imagine a large number of identical employers and a larger population of workers. Each employer will be randomly matched with many workers from this population. Workers belong to one of two identifiable groups, B or W. Denote by λ the fraction of W's in the population. The sole action of an employer is to assign each of his workers to one of two possible jobs, called tasks "zero" and "one." Task one is the more demanding and rewarding assignment.[6] While all workers can perform satisfactorily in task zero, a given worker may or may not be capable of satisfactory performance in task one.

All workers prefer to be assigned to task one, whether or not they are qualified (i.e., capable of satisfactory performance). Employers want to assign workers to task one only if they are qualified. Workers get the gross benefit ω if assigned to task one. Employers gain a net return $x_q > 0$ if they assign a qualified worker to task one and $-x_u < 0$ if they assign an unqualified worker. Define $r \equiv x_q / x_u$ to be the ratio of net gain to loss. Workers' gross returns and employers' net returns from an assignment to task zero are normalized to zero.[7]

Employers are unable to observe (prior to assignment) whether a worker is qualified for task one. Employers observe each worker's group identity and a noisy signal $\theta \in [0,1]$. The distribution of θ depends, in the same way for each group, on whether or not a worker is qualified. This signal might be the result of a test, an interview, or some form of on-the-job monitoring. Let $F_q(\theta)$ $[F_u(\theta)]$ be the probability that the signal does not exceed θ, given that a worker is qualified [unqualified] and let $f_q(\theta)$ and $f_u(\theta)$ be the related density functions. Define $\varphi(\theta) \equiv f_u(\theta)/f_q(\theta)$, to be the likelihood ratio at θ. We assume that $\varphi(\theta)$ is nonincreasing on $[0,1]$, which implies $F_q(\theta) \leq F_u(\theta)$ for all θ. Thus, higher values of the signal are more likely if the worker is qualified, and for a given prior, the posterior likelihood that a worker is qualified is larger if his signal takes a higher value.

Employers' assignment policies will be characterized by the choice of threshold "standards" for each group, such that only those workers with a signal observed to exceed the standard are assigned to the more demanding task. We will formalize this below, but intuitively what we have in mind is that employers are concerned about making two types of error in the classical statistical sense: assigning an unqualified worker, or failing to assign a qualified worker, to task one. Employers' beliefs about the likelihood that a worker is qualified will affect how they resolve this trade-off in the decision process. Since group membership is observ-

[6]This assignment can be thought of either as taking place at the time of matching or after the worker has spent a period of time in an entry-level position. In the latter interpretation, assignment to task one can be interpreted as promotion.

[7]The agents' payoffs represent the present value of all benefits to each party associated with a task-one assignment rather than a task-zero assignment. Wages are implicit in these payoffs and, given the task, are assumed to be equal for both groups. We treat wages as exogenous throughout the analysis. In particular, we will abstract from the possibility that wages change as a result of affirmative action. It is possible to extend the analysis, allowing an endogenous wage premium for task-one assignment, though the resulting model is much more complex. We do not believe that our results about the effect of affirmative action on employers' beliefs are sensitive to this assumption concerning wage determination.

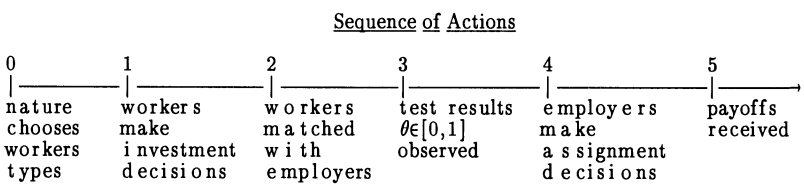

FIGURE 1. SEQUENCE OF ACTIONS

able, different beliefs about the likelihood that a group's members are qualified will lead to different standards for members of the groups. In this way, negative prior beliefs will bias the assignment process.

We assume that workers are qualified to do task one only if they have made some costly *ex ante* investment. This investment may be thought of either as acquiring knowledge (working hard at high school) or as acquiring life skills (developing good manners and work habits). The cost of becoming qualified varies among workers. Suppose for now that the cost distribution is the same for each group. Let c be a worker's investment cost and let $G(c)$ be the fraction of workers with investment cost no greater than c. Workers must decide, prior to being matched with an employer, whether making the investment is worthwhile. This depends on the extent to which investing raises the chance of being assigned to the more rewarding task, and hence on the standards the workers expect to face.

The timing of the interaction between workers and employers is summarized in Figure 1. First, nature chooses workers' "types," that is, their group membership (B or W) and their investment costs. Next workers decide whether or not to invest. Then they are matched with employers who, observing their group identities and signals, make assignment decisions. Employers' beliefs about the likelihood of a group's members being qualified will determine the standards they choose. These standards will, in turn, determine the fraction of each group who become qualified. Equilibrium is then a pair of employer beliefs which are self-confirming. A *discriminatory equilibrium* is one in which workers from one group (B's, say) are believed less likely to be qualified.

In order to define equilibrium formally, we must describe employers' and workers' behavior in more detail. We begin with employers' assignment decisions. Consider a worker belonging to a group, the representative member of which (according to an employer's prior beliefs) has probability $\pi \in (0,1)$ of being qualified. If that worker "emits" the signal θ then, using Bayes' Rule, the employer's posterior probability that he is qualified is the number $\xi(\pi,\theta)$ given by

$$(1)\quad \xi(\pi,\theta) \equiv \pi f_q(\theta)/[\pi f_q(\theta)+(1-\pi)f_u(\theta)] = 1/\{1+[(1-\pi)/\pi]\varphi(\theta)\}.$$

Having observed the worker's group and his signal, the employer's expected payoff from assigning him to task one is therefore $\xi(\pi,\theta)x_q-(1-\xi(\pi,\theta))x_u$. Since the payoff from assigning him to task zero is zero, the employer's best policy is to assign him to task one if and only if $x_q/x_u \geq [1-\xi(\pi,\theta)]/\xi(\pi,\theta)$, or equivalently, if and only if $r \geq [(1-\pi)/\pi]\varphi(\theta)$.

Given our monotone-likelihood-ratio assumption, the employer does best to choose a threshold value of the signal $s^*(\pi)$ (i.e., a standard) and to adopt the policy "assign a worker from a group whose representative member has prior probability π of being qualified to task one if and only if that worker's signal is no less than the standard $s^*(\pi)$," where[8]

$$(2)\quad s^*(\pi) \equiv \min\{\theta \in [0,1] | r \geq [(1-\pi)/\pi]\varphi(\theta)\}.$$

[8]This minimum may fail to exist when $\varphi(\theta)$ is not continuous. Then we define $s^*(\pi)$ by taking the infimum in (2). If the inequality fails for all $\theta \in [0,1]$, then the employer assigns all workers to task zero.

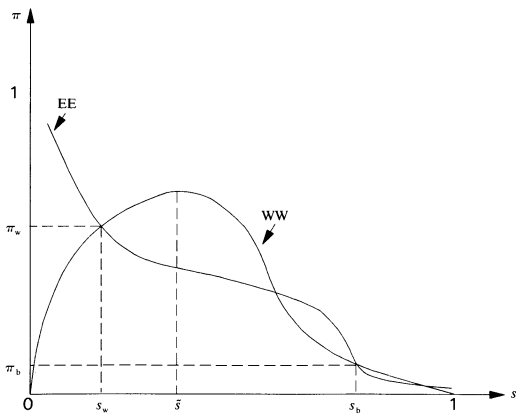

FIGURE 2. AN EQUILIBRIUM WITH NEGATIVE STEREOTYPES AGAINST B'S

Thus if, prior to observing any signal, employers believe that the probability is π_b (π_w) that a representative member of group B (W) is qualified, they will set the standards $s_i = s^*(\pi_i)$, $i = b$ and w. More optimistic beliefs about a group will be reflected in easier standards, since $s^*(\cdot)$ is decreasing in π.

We now turn to workers' investment decisions. The rational worker invests if the cost of doing so does not exceed the expected benefit. The expected benefit of investment is the product of two quantities: the gross return from being assigned to task one (ω) and the increased probability of assignment due to investing. The worker's assessment of the latter quantity depends on the standard he expects to face, since if the standard is s, the probability of assignment is $1 - F_q(s)$ when qualified, and $1 - F_u(s)$ when unqualified. Let $\beta(s) \equiv \omega[F_u(s) - F_q(s)]$ be the expected benefit of investment for any worker facing the standard s.

We conclude that a worker with investment cost c, facing the standard s, invests if and only if $c \leq \beta(s)$. Thus, among all workers facing the standard s, the proportion that become qualified is $G(\beta(s))$. The expected benefit $\beta(s)$ is a single-peaked function of s, increasing (decreasing) whenever $\varphi(s) > (<) 1$, and satisfying $\beta(0) = \beta(1) = 0$. These properties reflect the monotone-likelihood-ratio assumption, together with the fact that there is little point in investing when standards are very high or very low. Provided that G has a positive density over the relevant range and that $G(0) = 0$, it is also the case that $G(\beta(s))$ is single-peaked, rising (falling) with s as $\varphi(s) > (<) 1$, with $G(\beta(0)) = G(\beta(1)) = 0$.

A pair of beliefs for employers about the two groups will be *self-confirming* if, by choosing standards optimal for those beliefs, employers induce workers from the two groups to become qualified at precisely the rate postulated by the beliefs. Thus, we can define equilibrium as follows.

Definition 1: An *equilibrium* is a pair of beliefs (π_b, π_w) satisfying[9]

$$\pi_i = G(\beta(s^*(\pi_i))) \qquad i = b,w. \tag{3}$$

A discriminatory equilibrium (say, one with $\pi_b < \pi_w$) can occur whenever (3) has multiple solutions, for then it is possible that employers believe, consistent with their experience, that B's are less likely to be qualified than W's. Such discriminatory equilibrium beliefs reflect what we mean by "negative stereotypes." With these beliefs, employers force B's to meet a more exacting standard than W's in order to gain assignment to task one. This reduces the expected benefit from investment by B's, leading fewer of them to invest. In this way, the employers' initial negative beliefs are confirmed.

Figure 2 illustrates the analysis graphically. The horizontal axis measures the assignment standard(s), and the vertical axis

[9]Technically speaking, the interaction just described is a game of incomplete information with many players. In this game, nature chooses workers' types and matches workers with employers. A strategy for workers is a function $I(i,c)$ which gives a probability of investing for each worker type. A strategy for employers is a function $A(i,\theta)$ which gives the probability of assignment to task one for each state of information about a worker. An equilibrium is a strategy pair (I, A) such that each strategy is a best response to the other. It is easily verified that the self-confirming beliefs (π_b, π_w) of Definition 1 determine an equilibrium of this game in which workers and employers use the following strategies: $A(i,\theta) = 1$ (0) if $\theta \geq (<) s^*(\pi_i)$; and $I(i,c) = 1$ (0) if $c \leq (>) \beta(s^*(\pi_i))$.

measures the belief (π). The downward-sloping locus EE is the graph $\{(s,\pi)|s = s^*(\pi)\}$, depicting the standard-belief pairs consistent with optimal employer behavior. The hump-shaped curve WW is the graph $\{(s,\pi)|\pi = G(\beta(s))\}$, which represents pairs of standards and proportions of a group investing consistent with optimal worker behavior. The figure assumes $\varphi(\theta)$ to be smooth and strictly decreasing and assumes $G(c)$ to be continuous, with a positive density over the relevant range. If a point (s,π) lies on both curves then $s = s^*(\pi)$ and $\pi = G(\beta(s))$, so the belief π associated with that point solves (3).

Hence all equilibria can be identified in Figure 2 by associating each group with an intersection of the EE and WW curves. With multiple intersections, discriminatory equilibria exist. Note that $(s,\pi) = (1,0)$ solves (3) so long as $G(0) = 0$: the belief that no one in a group is qualified must be self-confirming, since it leads employers to assign everyone in that group to task zero, and no one would want to invest under those circumstances. Generally there are other equilibria, as is suggested by the following result.

PROPOSITION 1: *Assume that $\varphi(\theta)$ is continuous, strictly decreasing, and strictly positive on* $[0,1]$, *and that $G(c)$ is continuous and satisfies* $G(0) = 0$. *If there is an* $s \in (0,1)$ *for which* $G(\beta(s)) > \varphi(s)/[r + \varphi(s)]$, *then there exist at least two nonzero solutions of (3).*

PROOF:

Given the assumptions, EE lies above WW for s near 0 and 1, and both curves are continuous functions of s on $(0,1)$. Moreover, (2) implies that (s,π) is on the EE curve, $0 < s < 1$, if and only if

$$\pi = \varphi(s)/[r + \varphi(s)].$$

Therefore, $G(\beta) > \varphi/(r + \varphi)$ at s implies that EE lies below WW there. Hence the curves intersect at two or more distinct points where $\pi > 0$.

This proposition shows that statistical discrimination is a logically consistent notion in our model.[10] The existence of equilibria where employers hold negative self-confirming beliefs about some group does not require any assumptions about functional forms beyond those made in Proposition 1. Indeed, the sufficient condition given there must hold if either r or ω is large enough.

However, not all solutions of (3) are locally stable under the obvious adjustment process: $\pi^{t+1} = G(\beta(s^*(\pi^t)))$, $t = 0,1,2,\ldots$. This process converges to a solution π^* of (3), given an initial belief π^0 "close to" π^*, only if the absolute value of the slope of EE exceeds that of WW at π^*. A self-confirming belief that is not locally stable will not be robust to small errors of perception by employers and hence is less likely to be the basis of protracted discrimination against some disadvantaged group. Accordingly, it is important to identify whether or not particular equilibria are locally stable. In Figure 2, the solutions π_w, π_b, and zero are all locally stable in the above sense.[11]

Notice that stereotypes, in addition to being discriminatory, are also inefficient. When (3) has multiple solutions, the associated equilibria are Pareto rankable. To see this, let π_1 and π_2 be two self-confirming beliefs with $\pi_1 > \pi_2$. It follows that $s^*(\pi_1) < s^*(\pi_2)$. Hence, comparing π_1 with π_2, the following is true: workers are better off because they are more likely to be assigned to the more rewarding task, and employers are better off because they face a pool of more qualified workers. Thus we call the self-confirming belief π^* *Pareto efficient* if it is the largest solution of (3).

[10]Notice that in a discriminatory equilibrium employers' expected payoff from a W worker is higher than that from a B. We have ruled out the possibility of either W's being offered higher wages or employers refusing to hire B's. In effect, we are supposing that equal-pay laws prevent wage payments contingent on group identity and that fair-hiring laws prevent employers from simply refusing to deal with those B's with whom they have been randomly matched.

[11]Under the assumptions of Proposition 1, the no-investment equilibrium is locally stable. A little more structure is required to guarantee the existence of multiple locally stable equilibria.

When employers hold negative stereotypes they are not "color-blind." They correctly perceive group identity to be correlated with worker productivity, and they use this information to interpret the noisy signal. Since their beliefs are consistent with their experiences, they are acting rationally. However, as in Arrow's (1973) work, group identity conveys information only because employers expect it to.[12] If employers, or external observers, attribute the resultant inequality to inherent limitations of the less productive group, they are mistaken. This misattribution to an exogenous cause of what is in fact an endogenous difference seems to be an important feature of how stereotypes work in practice. *Webster's New World Dictionary* defines "stereotype" as "A fixed idea or popular conception about how a certain type of person looks, acts, etc." An agent with a "fixed idea" about a group, backed by evidence, may be unwilling to consider that his own and others' behavior is directly responsible for validating the generalizations upon which he acts.

However, an equilibrium with stereotypes does not require any such misattribution by employers.[13] Even if they all recognized the mechanism at work here, no single employer could reduce group productivity differences by altering his own assignment strategy. The action of a single employer will not affect investment incentives when workers do not know with which employer they will be matched. Breaking the negative stereotype requires that employers act in concert or that government somehow intervene. Affirmative-action policy, by forcing employers to assign workers about whom they have negative beliefs to task one more frequently, might be a useful instrument for this purpose. We investigate this possibility in the next two sections.

II. Affirmative Action

A. Extending the Basic Model

Let us consider now how a regulatory authority might intervene with some affirmative-action policy to break an equilibrium with stereotypes.[14] The simplest intervention would insist that employers make color-blind assignments, requiring that B's and W's with equal test scores be treated equally. This would create equivalent investment incentives for the two groups of workers, causing them to invest at the same rate and leading employers to revise their discriminatory beliefs. However, this policy can be enforced only if, in every instance, the regulator can observe all information upon which employers rely when making an assignment decision. Such a stringent informational requirement is unlikely to be met

[12] Related ideas can also be found, for example, in George Akerlof (1976), David Starrett (1976), and Andrew Weiss (1984).

[13] If one is willing to accept the possibility of such misattribution, the type of discrimination identified in this paper could easily arise in the interaction between a single employer and its workers. A worker's suitability for promotion is likely to depend not only on innate ability, but also on investment decisions made in his early years with the firm (learning "how things are done," establishing cordial relations with other employees, etc.). An employer who believes that minority workers have, on average, less innate ability (different investment cost distributions, say) may easily find his beliefs being confirmed in equilibrium through the type of mechanism identified here. An employer who fully understands the structure of the interaction, however, would experiment with different promotion standards to determine the validity of his beliefs.

[14] Intervention might not be necessary if the forces of competition could be relied upon to eliminate firms with negative stereotypes. This possibility is not considered in our model, since all employers are taken to have the same beliefs. In equilibrium, this homogeneity of beliefs is justifiable because employers are drawing from a common pool of workers and thus face statistically identical populations. Nevertheless, it would be interesting to consider how and whether such an equilibrium state would be reached if employers initially began with different beliefs and if the matching process associating workers and employers allowed for some element of self-selection. Even if there are forces that tend to undermine discriminatory beliefs in the long run, one still might find intervention of the sort we consider useful, since the government's actions could speed the transition process, especially if markets are less than perfectly competitive.

in practice.[15] Hence, we rule out the use of this policy, assuming in effect that in any worker–employer interaction the assignment outcome, but not the signal value θ, is observable (or verifiable) by an outside party. In this paper "affirmative action" refers to a policy requiring employers to achieve the same aggregate rate of assignment to task one for both groups. Our analysis applies most readily to those situations in which affirmative action takes mainly a "results-oriented" rather than a "process-oriented" form.[16]

The model is readily extended to incorporate this kind of regulation. Workers' behavior is not affected by the policy; they continue to make their investment decisions as before, depending on the assignment standards which employers use for each group. Thus, a group of workers' best-response behavior can still be represented by the WW curve. Affirmative action changes an employer's problem, however, because standards can no longer be chosen independently for the two groups. Rather, each employer must ensure that, whatever standards he uses, anticipated group rates of assignment to task one are equal.

Consider a group of workers about which an employer believes the fraction π are qualified, and for which he uses the assignment standard s. Let $\rho(s,\pi)$ be the probability the employer assesses to assigning a randomly drawn worker from this group to task one, and let $P(s,\pi)$ be the employer's expected payoff from such a worker. Then

$$\rho(s,\pi) \equiv \pi[1-F_q(s)] + (1-\pi)[1-F_u(s)]$$

and

$$P(s,\pi) \equiv \pi[1-F_q(s)]x_q - (1-\pi)[1-F_u(s)]x_u.$$

It follows that, under affirmative action, given beliefs (π_b,π_w), an employer will choose standards (s_b,s_w) to solve the following problem (where λ is the fraction of W's in the population):

$$(4)\quad \max[(1-\lambda)P(s_b,\pi_b) + \lambda P(s_w,\pi_w)]$$

$$\text{subject to } \rho(s_b,\pi_b) = \rho(s_w,\pi_w).$$

That is, an employer's best response to any pair of beliefs is to choose a pair of standards maximizing his expected payoff per worker, subject to the affirmative-action constraint.[17] This suggests the following definition of equilibrium in the presence of affirmative action:

Definition 2: An *equilibrium under affirmative action* is a pair of beliefs (π_b,π_w) and of standards (s_b,s_w) satisfying the following conditions:

(a) (s_b, s_w) solves problem (4), given (π_b,π_w);
(b) $\pi_i = G(\beta(s_i))$, $i \in \{b,w\}$.

[15] This point is also stressed by Lundberg (1991). For a graphic illustration of the difficulty, consider the problem an outsider would face in trying to judge whether the same standard has been employed in the making of two distinct tenure decisions.

[16] There has been considerable debate and uncertainty about precisely what firms must do to conform to affirmative-action guidelines. Chapter 2 of Nathan Glazer (1975) contains a dated but still useful discussion of the issues. Most affirmative-action programs involve some requirement that (in a suitable period of time) the representation of women and minorities in all positions be comparable to their availability in a pool of potential candidates, which accords with our modeling of the policy. Also, to the extent that a "process-oriented" program is undertaken in which the regulator has coarser information than the employer, enforcement of color-blind assignment behavior will have effects similar to those captured by the simple quantity constraint which we consider here.

[17] We are being somewhat casual here regarding how the government enforces its policy. Ideally one would like to leave employers' actions unrestricted, explicitly modeling their optimal response to whatever penalties are risked by violating the government's assignment guidelines. Instead, to keep things simple, we require all employers to set standards which they expect will cause the guidelines to be met, on the average. In the resulting setup an employer's feasible strategies [assignment policies satisfying the constraint in (4)] depend, in effect, on his beliefs. This is a departure from the usual formulation of a game with incomplete information.

Notice that $s^*(\pi)$ defined in (2) satisfies: $s^*(\pi) \in \text{argmax}\{P(s,\pi) | 0 \leq s \leq 1\}$.[18] Thus, the only difference between Definitions 1 and 2 is the addition of the requirement that $\rho(s_b, \pi_b) = \rho(s_w, \pi_w)$. However, if employers have homogeneous beliefs about the two groups, this constraint is not binding on their profit-maximizing choice of (s_b, s_w). Therefore, if π^* solves (3), then $\pi_b = \pi_w = \pi^*$ and $s_b = s_w = s^*(\pi^*)$ satisfy (a) and (b) of Definition 2. Therefore, if employers have the same beliefs about the two groups and, by using a common optimal standard, cause those beliefs to be confirmed, we have an equilibrium under affirmative action.

It is a highly desirable state of affairs that there exist no other equilibria under affirmative action. *When all equilibria under affirmative action entail homogeneous beliefs, a temporary color-conscious policy intervention by government must produce the permanent benefit of assuring employers' color-blind behavior.* Any preexisting negative stereotypes have to be eliminated. Moreover, once an equilibrium is reached, removal of the affirmative-action constraint will occasion no change in employers' behavior. It is therefore of some interest to determine circumstances under which affirmative-action policy necessarily produces this desirable outcome.[19]

A sufficient condition for this to be true is readily developed. Any group of workers facing the standard s invests so that the fraction $G(\beta(s))$ of them are qualified. Thus, if the standard for some group is s, in equilibrium employers must expect a fraction $\hat{\rho}(s) \equiv \rho(s, G(\beta(s)))$ of this group to be assigned to task one. Compliance with affirmative action makes employers equate $\rho(s,\pi)$ for both groups; but then self-confirming beliefs imply that $\rho(s,\pi) = \hat{\rho}(s)$ for each group. Thus, in any equilibrium under affirmative action, $\hat{\rho}(s_b) = \hat{\rho}(s_w)$. Now note that $\hat{\rho}(\cdot)$ must be decreasing over some part of its domain. After all, employers would expect to assign all workers to task one with a zero standard $[\hat{\rho}(0) = 1]$ and none with a standard of one $[\hat{\rho}(1) = 0]$. If $\hat{\rho}(\cdot)$ were decreasing over its entire domain, then s_b must equal s_w, and hence π_b must equal π_w. We have therefore established the following proposition.

PROPOSITION 2: *If $\hat{\rho}(\cdot)$ is decreasing on $[0,1]$, then all equilibria under affirmative action entail homogeneous beliefs about the two groups.*

How $\hat{\rho}(\cdot)$ varies with s depends on the interaction of two distinct effects. First, as s rises, access to task one is more strictly rationed; workers now need a higher test score to gain that assignment. This effect reduces the fraction of workers assigned to task one. Second, as s rises, the fraction of qualified workers changes. If s is smaller (larger) than $\tilde{s}$ in Figure 2 [defined by $\varphi(\tilde{s}) = 1$], increasing s raises (lowers) the fraction of investors. Obviously, the fraction of workers assigned to task one is increasing in the fraction of investors. Thus, while $\hat{\rho}(\cdot)$ is necessarily decreasing on $[\tilde{s}, 1]$, it may not be on $[0, \tilde{s})$. The positive investment effect may outweigh the stricter rationing effect.

Understanding intuitively when this will happen is difficult. The size of the stricter rationing effect depends on the properties of the particular testing technology. These properties, together with the distribution of investment costs and the payoff from being assigned to task one, also influence the magnitude of the investment effect. A simple calculation shows that $\hat{\rho}'(s) < 0$ on $[0,1]$ if and only if

$$(5) \qquad \varphi(s)/[\varphi(s) - 1] > \eta(\beta(s))$$

[18] Notice that

$$\partial P(s,\pi)/\partial s \gtreqless 0 \text{ as } r \gtreqless [(1-\pi)/\pi]\varphi(s).$$

Thus, the first-order condition for maximizing $P(s,\pi)$ with respect to s (allowing for the possibility of corner solutions) is satisfied by $s^*(\pi)$ defined in (2), and the second-order condition is guaranteed by the monotonicity of the likelihood ratio $\varphi(\theta)$.

[19] The term "desirable" should be interpreted with some care. Both groups may be made worse off as a result of the policy, despite the elimination of negative stereotypes. Thus, rather than improving employers' views of B's, the policy could lessen their opinion of W's. Were this to happen, the result would be Pareto inferior to the original situation.

for all $s \in [0, \tilde{s})$, where

$$\eta(c) \equiv d[c \cdot G(c)]/dc.$$

Now the left-hand side of (5) rises with s, as does $\beta(s)$ when $s < \tilde{s}$, so, if $\eta(c)$ is increasing on $[0, \beta(\tilde{s})]$, a sufficient condition for (5) is $\varphi(0)/[\varphi(0) - 1] > \eta(\beta(\tilde{s}))$, which must hold if $\varphi(0)$ is small enough and may hold when $\beta(\tilde{s})$ is small.

To illustrate, let costs be uniformly distributed on $[0, 2\mu]$; then $\eta(c) = c/\mu$, $0 \le c \le 2\mu$; so either $\varphi(0) < 2$ or $\beta(\tilde{s}) \le \mu$ implies (5). If costs are exponentially distributed with mean μ, then $\eta(c) = (c/\mu - 1) \times \exp[-c/\mu] + 1$, so $\eta(c)$ has its maximum at $c = 2\mu$, and $\eta(2\mu) = 1 + e^{-2}$. Thus, either $\varphi(0) < 1 + e^2 \approx 8.4$ or $\beta(\tilde{s}) \le \mu$ implies (5). Note that $\varphi(0)$ is a rough measure of the informativeness of the noisy signal; when $\varphi(0)$ is large, a low signal value is strong evidence that a worker did not invest. Moreover $\beta(\tilde{s})/\mu$ is the largest feasible investment benefit–cost ratio for the average worker. These illustrative examples therefore suggest the following rough rule of thumb. *Suppose that either (i) the noisy signal is relatively uninformative about workers' investment decisions or (ii) the cost distribution and payoffs are such that the average worker, even when facing maximal incentives, perceives acquiring the skill needed for task one to be a poor investment. Then affirmative action will eliminate stereotypes.*

The question which now arises is: what happens when the sufficient condition is not satisfied? To get some insight into this we will work through an example. A general treatment is provided in Section III, and the reader anxious to get to the main result can skip the example with no loss of continuity.

B. *Patronizing Equilibria in an Example with Uniform Distributions*

Consider a special case of this model in which the cost and signal distributions are assumed to be as follows: costs are uniform on $[0, 1]$; a qualified worker's signal is uniform on $[\theta_q, 1]$; an unqualified worker's signal is uniform on $[0, \theta_u]$; and $\theta_q < \theta_u$. In effect, there exists a test of qualification which yields one of three outcomes: "pass" ($\theta > \theta_u$); "fail" ($\theta < \theta_q$); and "unclear" ($\theta_q \le \theta \le \theta_u$). An employer is sure that a worker is (not) qualified whenever $\theta > \theta_u$ ($\theta < \theta_q$); and while the test is ambiguous when $\theta_q \le \theta \le \theta_u$, an employer has the same information for any such θ, because the likelihood ratio $\varphi \equiv (1 - \theta_q)/\theta_u$ is constant in this range. Let p_q (p_u) be the probability that, if a worker does (does not) invest, his test outcome is unclear. Then $p_q \equiv (\theta_u - \theta_q)/(1 - \theta_q)$, $p_u \equiv (\theta_u - \theta_q)/\theta_u$, and $\varphi = p_u/p_q$.

In the absence of affirmative action, an employer assigns "passers" to task one and "failers" to task zero. His decision in the event of an unclear test result depends on his beliefs. Let π be the employer's prior probability that a worker is qualified and let ξ be his posterior likelihood that the worker has invested given an unclear test result. Then Bayes' Rule implies that

$$\xi = 1/\{1 + [(1 - \pi)/\pi]\varphi\}.$$

The employer will assign the worker to task one only if $\xi x_q \ge (1 - \xi)x_u$. This is equivalent to $\pi \ge \varphi/(r + \varphi) \equiv \hat{\pi}$, so a worker with an unclear test gets the "benefit of the doubt" only if the employer is sufficiently optimistic about his group. An employer is "liberal" toward group i if he gives group-i workers the benefit of the doubt and "conservative" if he does not. A liberal policy amounts to choosing the standard $s = \theta_q$; a conservative one implies the standard $s = \theta_u$.

A worker's investment choice depends on how he anticipates employers will treat an unclear test result. If employers follow a liberal policy, a worker who has invested is assigned to task one for sure, while a noninvestor is assigned with probability p_u. Thus the expected benefit from investing is $\pi_\ell \equiv \omega(1 - p_u)$. When employers are conservative, a noninvestor will have no chance of being assigned to task one, while an investor will be assigned with probability $1 - p_q$. Thus the expected benefit from investing is $\pi_c \equiv \omega(1 - p_q)$. Since costs are uniformly distributed on $[0, 1]$, π_c (π_ℓ) is also the fraction of workers in a group who are qualified,

given the anticipated conservative (liberal) behavior of employers.

We conclude that π_ℓ (π_c) is a self-confirming belief if and only if $\pi_\ell \geq \hat{\pi}$ ($\pi_c < \hat{\pi}$). When $\pi_\ell \geq \hat{\pi}$, workers expecting to face liberal employers invest in sufficient numbers that being liberal is optimal for employers. When $\pi_c < \hat{\pi}$, workers expecting to meet conservative employers invest so infrequently that being conservative is an optimal employer response. Thus, in either case, were employers to hold the indicated belief, they would act in such a way that this belief would be confirmed by their experience. Therefore, in the absence of affirmative action, when $\pi_c < \hat{\pi} < \pi_\ell$ an equilibrium exists in which employers harbor negative stereotypes against B's: $(\pi_b, \pi_w) = (\pi_c, \pi_\ell)$. Here employers are pessimistic about and conservative toward B's, while being optimistic about and liberal toward W's. A sufficient condition for this equilibrium to exist is

$$(6)\quad \omega(1-p_q) < x_u p_u / [x_q p_q + x_u p_u] < \omega(1-p_u).$$

This equilibrium is locally stable, since small changes in beliefs do not cause employers to revise their standards.

Assume that (6) is satisfied and that we are in such a discriminatory equilibrium. What would be the effect of introducing affirmative action? Costs are distributed uniformly on [0,1], so that, by our earlier argument, either $\varphi(0) < 2$ or $\beta(s) \leq \frac{1}{2}$, $0 \leq s \leq 1$, would guarantee that (5) holds; but the signaling distributions in the example imply $\varphi(0) = +\infty$. Also, (6) implies that investment incentives are maximal when the employer is liberal ($\tilde{s} = \theta_q$). Therefore, if $\beta(\theta_q) = \omega(1-p_u) = \pi_\ell > \frac{1}{2}$, we cannot use the analysis above to ensure that affirmative action produces benign results in this example. Indeed, quite to the contrary, we can establish the following dramatic result.

PROPOSITION 3: *Assume that* $\pi_\ell > \hat{\pi} > \pi_c$, $\pi_\ell > \frac{1}{2}$, *and* $\lambda < 1$ *is sufficiently large. Then in the only stable equilibrium under affirmative action, given the obvious adjustment process, employers continue to hold negative stereotypes about B's. In fact, their (correct) assessment of the average productivity of B's may actually worsen in this equilibrium.*

The basic logic of this result is simple: to comply with an equal-assignment mandate, and believing B's to be less productive, employers *patronize* B's by making it easier for them to achieve the desirable assignment. This is optimal for employers when B's are relatively few in the population. However, because it is easier for them to succeed, B's find it less profitable to invest, thus confirming employers' negative views. This causal chain has the interesting feature that, though B's face a lower standard than W's, they respond to it in such a way that they end up assigned to task one at the same rate as W's. Thus, the effect on B's of less severe rationing is just offset by the reduced investment incentives of a lower standard. This is precisely what (5) rules out.

To establish the proposition, we begin by noting that compliance with the mandate of affirmative action requires that more B's or less W's be assigned to task one. Given any beliefs for which $\pi_b < \pi_w$, it should be intuitively clear that, if B's are rare enough in the population (i.e., if λ is large enough), compliance is best achieved by increasing the rate at which B's are assigned to task one, not by lowering the rate for W's.

Indeed, when $\hat{\pi} < \pi_\ell = \pi_w$ there exists $\hat{\lambda} < 1$ such that, for $\lambda > \hat{\lambda}$ and any $\pi_b < \pi_\ell$, employers prefer to achieve compliance by assigning failing B's to task one than by assigning unclear W's to task zero.[20] Sup-

[20]Consider assigning either ΔB more B's to task one, or alternatively ΔW more W's to task zero, with the object in each case to reduce the difference in assignment rates to task one by the same amount. Then, $\Delta B/(1-\lambda) = \Delta W/\lambda$. At the initial equilibrium, an employer loses $\xi_\ell x_q - (1-\xi_\ell)x_u$ if he assigns an unclear W to task zero, while he loses x_u if he assigns a failing B to task one, where

$$\xi_\ell = \pi_\ell p_q / [\pi_\ell p_q + (1-\pi_\ell) p_u]$$

pose then that $\lambda > \hat{\lambda}$. Then, given *any* beliefs (π_b, π_w) such that $0 \le \pi_b \le \pi_w = \pi_\ell$, an employer's optimal solution to problem (4) involves assigning W's as before, assigning unclear B's to task one, and assigning failing B's to task one with a probability just large enough to achieve compliance. Let $\alpha(\pi_b)$ denote this probability. Then $\alpha(\pi_b)$ is defined by the equation

$$(7)\quad \pi_\ell + (1-\pi_\ell)p_u = \pi_b + (1-\pi_b)[p_u + (1-p_u)\alpha(\pi_b)]$$

which implies: $\alpha(\pi_b) = (\pi_\ell - \pi_b)/(1-\pi_b)$. Whenever an employer assigns a failing worker to task one, we say the employer is patronizing that worker.

Consider now workers' best response to this employer behavior. W's continue to invest at rate π_ℓ, since their incentives are unchanged. If a B worker expects to be patronized with probability α, his return from investing is $\omega(1-\alpha)(1-p_u)$, since the only way he can be assigned to task zero when he does not invest is that he fails the test and is not patronized, which occurs with probability $(1-\alpha)(1-p_u)$. Therefore, if B's anticipate being patronized with probability α, the fraction of them who invest is $\omega(1-\alpha)(1-p_u) = (1-\alpha)\pi_\ell$.

It follows that the beliefs (π_b, π_ℓ) can arise in an equilibrium of this example under affirmative action if and only if $\pi_b \le \pi_\ell$ and

$$(8)\quad \pi_b = [1-\alpha(\pi_b)]\pi_\ell = [(1-\pi_\ell)/(1-\pi_b)]\pi_\ell.$$

Since $\pi_\ell > \frac{1}{2}$, there are two possible equilibrium beliefs about B's: $\pi_b = \pi_\ell$, and $\pi_b = 1-\pi_\ell$. The former is the color-blind outcome, in which employers are liberal toward both groups. Unfortunately, the only stable equilibrium is the patronizing one, $\pi_b = 1-\pi_\ell < \pi_\ell = \pi_w$, where employers continue to see B's as less productive.

To see this, note that if employers start with beliefs $(\pi_b^0, \pi_w^0) = (\pi_c, \pi_\ell)$ then, in view of the foregoing discussion culminating in (8), at stage t of the obvious adjustment process their beliefs are (π_b^t, π_ℓ), where $\{\pi_b^t\}$ solves the following difference equation:

$$(9)\quad \pi_b^{t+1} = [1-\alpha(\pi_b^t)]\pi_\ell = [(1-\pi_\ell)/(1-\pi_b^t)]\pi_\ell \qquad t = 0,1,2,\ldots;\ \pi_b^0 = \pi_c.$$

The reader can easily verify that for $\pi_\ell > \frac{1}{2}$ the solution of (9) converges to $1-\pi_\ell$ as $t \to \infty$. Thus, the only stable equilibrium is the patronizing one. Note that if $\pi_\ell + \pi_c > 1$, the stereotype against B's worsens under affirmative action $(\pi_b = 1-\pi_\ell < \pi_c = \pi_b^0)$.[21] This occurs if ω is large (a big benefit–cost ratio for the average worker) or if p_u and p_q are small (a highly accurate test). Even if beliefs about B's are not worsened, when $\pi_c \le 1-\pi_\ell < \hat{\pi}$ affirmative action will have to be a permanent fixture for B's gains to continue, since otherwise employers revert to conservative behavior toward B's as soon as the constraint is removed.

The reader may suspect that this counterintuitive outcome depends in some way on the special features of this example—notably, the fact that the likelihood ratio $\varphi(\theta)$ is not bounded, continuous, or strictly positive on $[0,1]$. However, as we show in the next section, patronization can occur when all the distribution functions are smooth, for a nonnegligible range of parameter values.

III. The Main Result

To pursue the analysis further we must consider problem (4) in more detail. The Lagrangian for the employers' constrained optimization problem can be written as

so he would rather put failing B's into task one than put unclear W's into task zero, to narrow the gap by a given amount, if $[\lambda/(1-\lambda)][\xi_\ell x_q - (1-\xi_\ell)x_u] > x_u$ [i.e., if $\lambda > \hat{\lambda} \equiv 1/\xi_\ell(1+r)$]. Note that $\pi_\ell > \hat{\pi}$ implies $\hat{\lambda} < 1$.

[21] The reader may find it helpful to experiment with some numerical examples. Suppose, for example, that $p_u = 0.2$, $p_q = 0.3$, and $r = \frac{2}{3}$. Then if $\lambda > 0.9$, for values of ω such that $\frac{5}{7} > \omega > 0.5/[\lambda - 0.2]$, patronization of B's is the result of affirmative action. The negative stereotype about B's is made worse if, in addition, $\omega > \frac{2}{3}$.

follows:

$$(10)\quad \mathcal{L}(s_b, s_w, \gamma; \pi_b, \pi_w) = [(1-\lambda)P(s_b, \pi_b) + \lambda P(s_w, \pi_w)] + \gamma[\rho(s_b, \pi_b) - \rho(s_w, \pi_w)]$$

where γ is a multiplier associated with the affirmative-action constraint. Suppose that the functions $P(\cdot,\cdot)$ and $\rho(\cdot,\cdot)$ are continuously differentiable, and that $\varphi(\cdot)$ is decreasing. Then an interior solution is fully characterized by the first-order conditions: $\partial\mathcal{L}/\partial s_i = 0$, $i \in \{b,w\}$, and $\partial\mathcal{L}/\partial\gamma = 0$.[22] By the Kuhn-Tucker theorem, for given beliefs (π_b, π_w), any triple (s_b, s_w, γ) satisfying these three conditions identifies a solution of the employer's problem (4). These beliefs and associated optimal standards are an equilibrium in the sense of Definition 2 if, in addition, $\pi_i = G(\beta(s_i))$, $i \in \{b,w\}$. Notice that the multiplier γ must be positive (zero) when $\pi_b < \pi_w$ ($\pi_b = \pi_w$).

Suppose then that $\pi_b \le \pi_w$ and, for arbitrary $\gamma \ge 0$, consider the first-order conditions $\partial\mathcal{L}/\partial s_i = 0$, $i \in \{b,w\}$. After some manipulation, these conditions may be expressed as follows:

$$(11a)\quad r_w(\gamma) \equiv [x_q - \gamma/\lambda]/[x_u + \gamma/\lambda] = [(1-\pi_w)/\pi_w]\varphi(s_w)$$

and .

$$(11b)\quad r_b(\gamma) \equiv [x_q + \gamma/(1-\lambda)]/[x_u - \gamma/(1-\lambda)] = [(1-\pi_b)/\pi_b]\varphi(s_b).$$

These equations, contrasted with (2), have an instructive interpretation. Given a "shadow price of equality," $\gamma \ge 0$, employers act as if they must pay the tax γ/λ for each W assigned to task one instead of task zero, while receiving the subsidy $\gamma/(1-\lambda)$ for each B put into task one rather than task zero. Therefore, employers generally respond to the affirmative-action constraint by lowering the assignment standard for B's and raising it for W's; and these adjustments are larger for B's, and smaller for W's, the larger is λ.

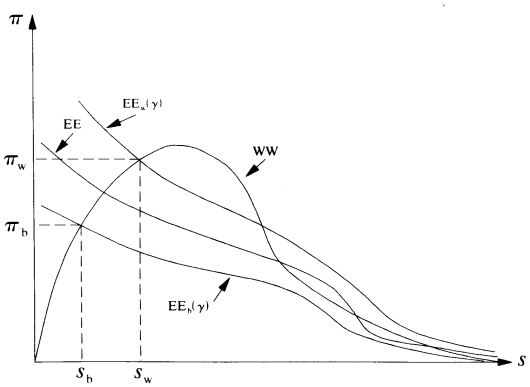

FIGURE 3. EMPLOYERS' OPTIMAL STANDARDS FOR B'S AND W'S GIVEN BELIEFS AND POSITIVE VALUE OF MULTIPLIER

Equations (11) allow us to extend the graphical analysis of Figure 2 so as to study equilibria under affirmative action. Given $\gamma \ge 0$, (11) defines two graphs in the (s,π) plane which we call the $EE_w(\gamma)$ and $EE_b(\gamma)$ curves, respectively. These curves are depicted in Figure 3. For any beliefs (π_b, π_w), and any multiplier γ, standards satisfying first-order conditions (11) are found at points (s_i, π_i) on the $EE_i(\gamma)$ curves, $i \in \{b,w\}$.

Now consider in Figure 3 the intersections of these $EE_i(\gamma)$ loci with the WW curve which, as before, is the graph $\{(s,\pi) | \pi = G(\beta(s))\}$. The standards and beliefs at these two points satisfy (11) for this value of γ and also have the property that the beliefs would be self-confirming were employers to adopt those standards. Thus these two points depict an equilibrium in the sense of Definition 2 if, in addition, they satisfy the affirmative-action constraint. Figure 4 extends the diagram to include this constraint. Figure 4A exhibits $\hat{\rho}(s)$, and

[22] Second-order conditions are guaranteed since problem (4) is quasi-concave, in view of the monotone-likelihood-ratio assumption. To verify this, set up the standard bordered Hessian matrix, use the fact that the cross-partial derivatives $\partial^2\mathcal{L}/\partial s_w \partial s_b \equiv 0$, and note that the principal minors of the Hessian alternate in sign, as required, when $\varphi'(s) < 0$.

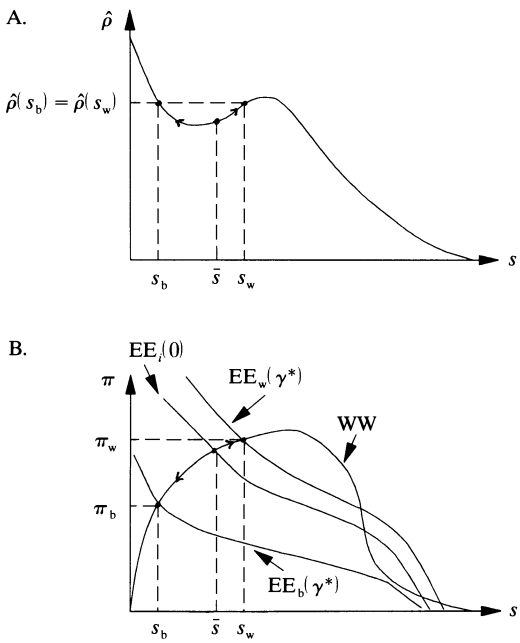

FIGURE 4. AN EQUILIBRIUM UNDER AFFIRMATIVE ACTION WITH NEGATIVE STEREOTYPE ABOUT B'S

Figure 4B shows the WW curve and the $EE_i(\gamma)$ loci, for $\gamma = 0$ and for some $\gamma^* > 0$. The two curves coincide when $\gamma = 0$. As γ grows larger, the implicit subsidy to B's and tax on W's in task one increases, so the $EE_b(\gamma)$ curve shifts down, and the $EE_w(\gamma)$ curve shifts up.

Figure 4 is so constructed that, for the particular multiplier value $\gamma^* > 0$, the affirmative-action constraint is satisfied at the indicated intersections of the $EE_i(\gamma^*)$ curves with the WW curve, $i \in \{b, w\}$. Thus, *these points depict an equilibrium under affirmative action in which employers have negative stereotypes about B's*. The question is whether there exists a multiplier $\gamma^* > 0$ for which the situation illustrated in Figure 4 actually obtains. Our main result provides the answer to this question.

PROPOSITION 4: *Assume F_u and F_q are continuously differentiable on $[0,1]$, G is continuously differentiable on $[0, \beta(\hat{s})]$, $\varphi'(\theta) < 0$ on $[0,1]$, and $G(0) = 0$. Suppose $\hat{\rho}'(s) > 0$ for some $\bar{s} \in (0, \tilde{s})$. Then there is a nonempty open set of parameters (λ, ω, r) such that, for any of these parameters, an equilibrium under affirmative action exists exhibiting negative stereotypes toward B's. Moreover, if $\lambda' > \lambda$ then such an equilibrium also exists for (λ', ω, r).*

PROOF:

Consult Figure 4. For $\bar{s} \in (0, \tilde{s})$ with $\hat{\rho}'(\bar{s}) > 0$, choose $(\bar{\omega}, \bar{r})$ such that the $EE_i(0)$ and WW loci intersect at $s = \bar{s}$ (i.e., the parameters $(\bar{\omega}, \bar{r})$ satisfy $\bar{r} = \varphi(\bar{s})[1 - G(\beta)]/G(\beta)$, where $\beta \equiv \bar{\omega}[F_u(\bar{s}) - F_q(\bar{s})]$). We will show that for any such $(\bar{\omega}, \bar{r})$, and for $\lambda < 1$ but sufficiently large, there is a multiplier $\gamma^*(\bar{\omega}, \bar{r}, \lambda) > 0$ such that the intersections of the $EE_i(\gamma^*)$ curves with the WW curve shown in Figure 4B, $i \in \{b, w\}$, have the property $\hat{\rho}(s_b) = \hat{\rho}(s_w)$.

Consider how the figure changes as γ rises from zero. As the EE_b curve shifts down and the EE_w curve shifts up, they trace out intersections with the WW curve. Denote by $s_i(\gamma)$ the value of s at the intersection of the $EE_i(\gamma)$ curve with the WW curve in the neighborhood of $\bar{s}$. The standards $s_i(\gamma)$ satisfy equations (11), and $s_i(0) = \bar{s}$, for $i \in \{b, w\}$. Applying the implicit-function theorem to (11) permits us to take $s_i(\cdot)$ as differentiable functions in a neighborhood of zero whose radius depends on λ. It is clear that $s_b'(\gamma) < 0$ and $s_w'(\gamma) > 0$. Also, since $G(\beta(s)) \to 0$ as $s \to 0$, it follows from (11b) that $s_b(\cdot)$ varies continuously with γ for $\gamma \in [0, (1-\lambda)x_u)$ and that $s_b(\gamma) \to 0$ as $\gamma \to (1-\lambda)x_u$. Moreover, (11a) implies that the region where $s_w(\gamma)$ varies continuously with γ is larger, the larger is λ.

Combining these observations we conclude that when λ is sufficiently close to 1, as γ rises from 0 to $(1-\lambda)x_u$, $s_b(\gamma)$ falls smoothly from $\bar{s}$ toward 0, and $s_w(\gamma)$ rises smoothly from $\bar{s}$. Now, let $D(\gamma) \equiv \hat{\rho}(s_b(\gamma)) - \hat{\rho}(s_w(\gamma))$. $D(\cdot)$ is differentiable for γ near 0, and $D'(0) < 0$; and, since $\hat{\rho}(s_b(\gamma)) \to 1$ as $\gamma \to (1-\lambda)x_u$, $D(\gamma') > 0$ for some $\gamma' \in (0, (1-\lambda)x_u)$. Thus, there is a $\gamma^* \in (0, \gamma')$ at which $D(\gamma^*) = 0$. Hence, an equilibrium under affirmative action with negative stereotypes against B's exists for parameter values $(\bar{\omega}, \bar{r}, \lambda)$ if λ is large enough. This

conclusion can be seen graphically as well, in Figure 4A. For λ near 1, as γ rises from 0 the point $(s_b(\gamma), \hat{\rho}(s_b(\gamma)))$ "moves" down the graph of $\hat{\rho}$ away from $(\bar{s}, \hat{\rho}(\bar{s}))$ much faster than $(s_w(\gamma), \hat{\rho}(s_w(\gamma)))$ "moves" up the graph. Thus, eventually a positive value of the multiplier γ^* must be reached at which $\hat{\rho}(s_b(\gamma^*)) = \hat{\rho}(s_w(\gamma^*))$. To complete the proof notice that, given the continuity assumed, the qualitative features of Figure 4 will be unchanged for payoff parameters (ω, r) that are near $(\bar{\omega}, \bar{r})$.

Generalizing the terminology of Subsection II-B we call it a *patronizing equilibrium under affirmative action* if employers have (correct) beliefs about the inferiority of B's and therefore use a lower standard in order to be sure that B's are assigned to task one at the same rate as W's. The term "patronizing" is apt because, in an effort to assure B's success but believing them to be less capable than W's, employers treat B's more liberally, thereby ensuring that their negative beliefs become a self-fulfilling prophecy.

Whether affirmative action leads to an improvement in the perception of the capabilities of B's, relative to laissez-faire, depends on the circumstances. It is possible that, starting in a situation where employers are unconstrained and hold negative stereotypes about B's, the introduction of affirmative action, though leading to patronization, might raise employers' estimate of the productivity of B's by enough that, upon removal of the policy, beliefs about both groups would converge to the same (locally stable) equilibrium. However, as the example above showed, this need not be the case. In any event, when patronizing equilibria exist, a regulator cannot be sure that an intervention aimed at eradicating the use of group identity as a basis for occupational assignment will not instead have the unintended effect of encouraging the ongoing color-conscious behavior of employers.

IV. Further Policy Considerations

The major insight of this paper is that an equal-assignment constraint creates incentives for employers to make job-assignment decisions that interact in interesting and unexpected ways with the incentives workers have for acquiring skills. If employers begin believing that B's are inferior to W's $(\pi_b < \pi_w)$ they will be more conservative about assigning B's to demanding jobs. If with these same beliefs they are forced to assign those jobs to both groups at an equal rate, then they will switch to treating B's more liberally. Though the initial conservative treatment discouraged some B's from investing, the switch to treating B's more liberally than W's can also reduce their relative incentive to invest.

In particular, whenever s_b is less than $\bar{s}$ in Figure 2, B investment is discouraged by the use of a marginally more liberal standard. If employers' initial beliefs about W's are such that their ideal standard $s_w = s^*(\pi_w)$ is less than $\bar{s}$, and if B's are a relatively small fraction of the population, then the optimal employer response to the affirmative-action constraint is to leave s_w essentially unchanged while lowering s_b enough to achieve equal proportionate representation of both groups in task one. Proposition 4 shows that this behavior will be consistent with the requirement that beliefs be self-confirming as long as $\hat{\rho}'(s_w) > 0$. This is the logic of patronization in the general case.

This logic has significant implications for policy beyond those noted above. First, it implies that a modest program of affirmative action can have unintended negative effects, even when there is no negative stereotype against B's. This occurs when job preferences are used to reduce group disparities that arise out of *ex ante* inequality in the distribution of skills. To illustrate, suppose that, because of unequal educational opportunities (say), B's have higher investment costs than W's on average. Concretely, assume $G_b(c) < G_w(c)$ for $0 < c \leq \beta(\bar{s})$. Let $\hat{\rho}_i(s) \equiv \rho(s, G_i(\beta(s)))$ $(i = \{b, w\})$, and assume that $\hat{\rho}_i(\cdot)$ is decreasing for both groups. Thus, by Proposition 2, we know that the kind of patronization identified in Proposition 4 could not occur here.

Figure 5 depicts this situation. It modifies Figures 2 and 3, allowing a separate WW

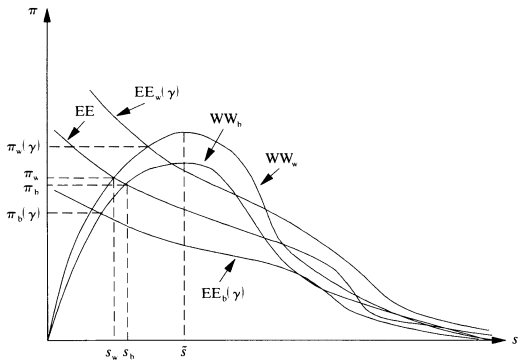

FIGURE 5. AFFIRMATIVE ACTION INCREASES SKILL DISPARITY IN THE ABSENCE OF STEREOTYPES

curve for each group, with WW_b lying below WW_w at each $s \in (0,1)$. Ignoring stereotypes, we focus on the two (Pareto efficient) self-confirming beliefs π_b and π_w depicted in the figure. B's are doing less well than W's, but the difference derives solely from their inferior endowments. Now consider the effect of a "marginal" affirmative-action policy. By this we mean a policy requiring a modest narrowing of the gap $\rho(s_w, \pi_w) - \rho(s_b, \pi_b)$, though not necessarily equal proportionate representation of the groups in task one.

Let $\gamma > 0$ be the multiplier on this constraint in an employer's profit-maximization problem analogous to (4). If the policy is moderate, γ will be small. Following the analysis of Section III, we see that introduction of the constraint shifts the EE curve up for W's and down for B's. Under the assumptions above this must increase the fraction of B's going to task one, reduce the fraction of W's, and so narrow the gap. Yet, in view of the fact that initially both $s_b < \tilde{s}$ and $s_w < \tilde{s}$, this marginal policy of affirmative action must also have the effect of exacerbating the difference $\pi_w - \pi_b$. That is, *using preferences to help the disadvantaged group necessarily causes the objective difference in productivity between the two groups to rise*. On the other hand, it is easy to verify that if the initial equilibria for both groups were in the range $(\tilde{s}, 1)$, then a marginal policy of job preferences for B's would also have had the effect of narrowing the (correctly) perceived disparity in group productivities, even as it raised the fraction of B's holding good jobs.

A second implication of the ambiguous incentive effects of employer-mediated group preferences is the fact that policies aimed directly at encouraging workers to invest generally avoid the pitfalls associated with affirmative action. At the same time, efforts to "bribe" employers to favor members of a particular group (instead of coercing them) are hampered by the same negative unintended consequences that can emerge with job quotas. To make this point we will compare the effects of two policies other than affirmative action which might be used to break an initial equilibrium with negative stereotypes: a subsidy to employers for placing B's in task one, and a subsidy to each B for getting assigned to task one by his employer. Both of these policies are feasible for a regulator having no more information than is required to enforce affirmative action, since they involve payments contingent only on assignment outcomes;[23] but these two policies have effects which differ from those induced by affirmative action, and from each other.

This is illustrated in Figures 6 and 7, which revert to the assumption of a common cost distribution for the two groups. Figure 6 envisions that employers are paid a subsidy of τ for each B assigned to task one. Figure 7 imagines that B's receive the payment τ, over and above their gross payoff ω, for being assigned to task one by

[23]Of course if the regulator could directly subsidize investment by B workers, the discriminatory equilibrium would be easily broken. However, such a subsidy would require that B workers' investment decisions be observable to the regulator, when we have assumed them to be unobservable to employers. We rule this out, since we are thinking of investment/effort decisions (like how hard one studies in school) which cannot be readily monitored. Indeed, overall efficiency could be improved through investment subsidies to both groups, because of the informational externality present here. The marginal investor does not consider that, by increasing the fraction of investors, employers would be induced to lower standards, thereby benefiting all workers.

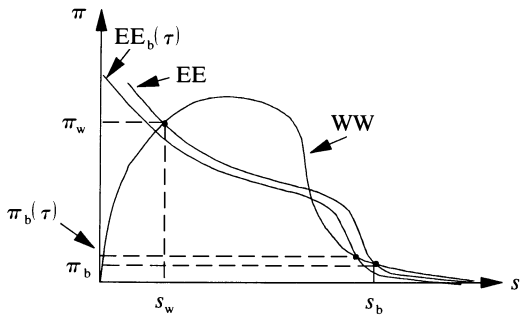

FIGURE 6. MARGINAL EMPLOYER SUBSIDY

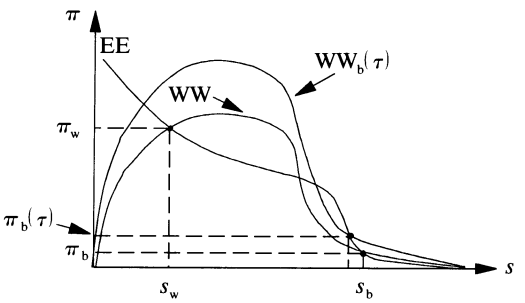

FIGURE 7. MARGINAL WORKER SUBSIDY

their employers. The employer subsidy raises their effective payoff ratio for B's from $r = x_q/x_u$ to $r_b \equiv (x_q + \tau)/(x_u - \tau)$, so it shifts down the EE curve applicable to B's. The worker subsidy raises their return from investing by the amount $\tau[F_u(s) - F_q(s)]$ at each standard s, thus shifting up the WW curve applicable to B's. (We rule out deals between employers and B's involving side payments, assuming that they would be unenforceable in court.) Notice that these group-B-specific subsidies will have no effect on the interactions between employers and W's.

Suppose initially that there is a discriminatory equilibrium with $0 < \pi_b < \pi_w$, and that a subsidy policy is enacted with the intent of breaking the negative stereotype against B's. Assume that both π_b and π_w are locally stable solutions of (3), so the EE curve cuts the WW curve from above at both points, and let the belief that employers hold about W's, π_w, be Pareto efficient. Now consider the effect of a "marginal" subsidy, one where τ is so small that the qualitative behavior of the set of self-confirming beliefs is unchanged.[24]

It is obvious from Figures 6 and 7 that such a subsidy, whether directed to employers or to workers, must reduce the difference in employers' beliefs about the productivity of B's and W's. This is because, whether EE shifts down or WW shifts up, the change implies a rise in π_b as long as the initial belief is nonzero, locally stable, and lies on the downward-sloping part of the WW curve. This last requirement must hold if employers initially held negative stereotypes toward B's, since EE and WW can intersect at most once on the upward-sloping part of WW. A marginal subsidy helps B's by setting in motion a mutually reinforcing process in which workers invest more when facing a lower standard and employers use lower standards when seeing evidence of greater investment.

However, it is also obvious that no marginal subsidy can ever completely eliminate the stereotype against B's. Such a policy produces a local improvement only; once it is removed, employers' beliefs [under the adjustment process $\pi^{t+1} = G(\beta(s^*(\pi^t)))$] eventually revert to what they had been in the original equilibrium.[25] To break the stereotype the subsidy must be "large"; but now the effect of subsidizing employers is quite different from that of subsidizing workers. Indeed, if employers' belief about W's lies on the upward-sloping part of the WW curve, there is no subsidy to employers for the assignment of B's to task one which can induce a revision of beliefs that eliminates the stereotype. Figure 8 shows that if the employer subsidy is large enough it can result in a more pessimistic view of B's than at the initial equilibrium. In this case the

[24]That is, τ is small enough that the set of solutions of (3), modified to allow for a subsidy of size τ', varies continuously as a function of τ', for $\tau' \in [0, \tau]$.

[25]This is because, by definition, a marginal subsidy cannot shift any solution of (3) outside of the "basin of attraction" of the original, locally stable, self-confirming belief.

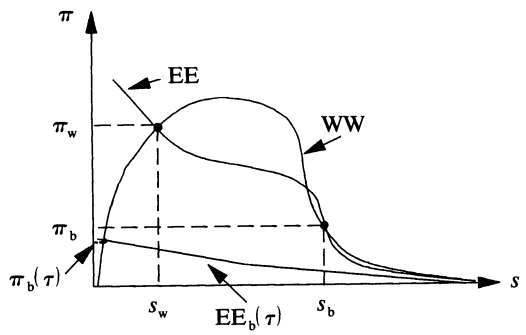

FIGURE 8. LARGE EMPLOYER SUBSIDY

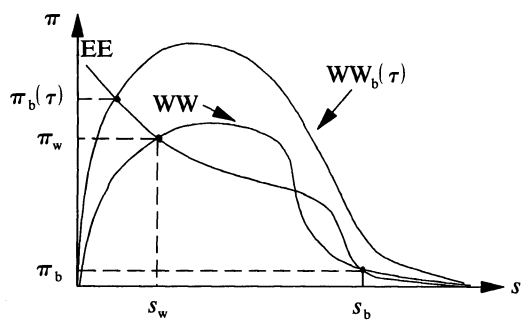

FIGURE 9. LARGE WORKER SUBSIDY

subsidy program backfires. It induces employers to lower their standards for B's so significantly that investment becomes less profitable, much as in patronizing equilibria under affirmative action.

A subsidy directed at B's does not have this problem, however. Although the equilibrium effect of a worker subsidy will always have employers using a lower standard, this must be accompanied by greater worker investment. A sufficiently large worker subsidy will overcome the stereotype by eliminating all locally stable nonzero self-confirming beliefs except the one on the upward-sloping part of the WW curve, shown in Figure 9, at which employers now believe B's to be superior to W's. A regulator could break the negative stereotype by imposing such a subsidy and then gradually phasing it out, arriving at a nondiscriminatory, Pareto efficient equilibrium.

Thus, we conclude that, generally speaking, it is better to subsidize disadvantaged workers for achieving good jobs, than to subsidize employers for promoting them, if the objective is to dispel negative self-confirming stereotypes.[26] A subsidy to workers increases their performance, no matter what employers' standards. A subsidy to employers causes them to lower their standards, which can also lower workers' performance, exacerbating the problem of negative stereotypic beliefs. As demonstrated in Section III, affirmative action has some of the same negative features identified here for employer subsidies.

There is, however, one important exception to this rule. When employers' views about B's are so negative that they assign none of them to task one ($\pi_b = 0$, $s_b = 1$), no subsidy to B's can break the discriminatory equilibrium. Since initially B's think the probability of assignment to task one is zero, none of them will incur the cost of investing, no matter how large the promised reward for achieving task one. Neither will a subsidy to employers be effective. If $\tau < x_u$, then employers, believing no B's are investing, will refuse to put any of them in task one; while if $\tau > x_u$ employers would want to assign all B's to task one, but then none of them will invest. In this situation, therefore, a policy of affirmative action would seem to be the only way to make progress.

V. Conclusion

A significant part of the debate over the desirability of affirmative action has focused on whether it can eliminate employers' negative stereotypes about the capabilities of minority workers. The key policy question underlying this concern is whether labor-market gains to minorities stemming from affirmative action can continue without it becoming a permanent fixture. This paper provides a theoretical analysis of this prob-

[26] In a standard supply–demand framework, the net effect of a specific subsidy is independent of whether it is paid to employers or to workers. This result does not emerge here because, given equal-pay laws, wages in a given task are constrained to be the same for both groups of workers.

lem. Using the idea of self-confirming discriminatory beliefs, we have formally analyzed a question which heretofore has resisted rigorous study: how will affirmative action affect stereotypes about minority workers?

The results of our study give credence to both the hopes of advocates of preferential policies and the concerns of critics. There are circumstances under which affirmative action will *necessarily* eliminate negative stereotypes. However, there are equally plausible circumstances under which it will not only fail to eliminate stereotypes, but may worsen them. This occurs because job preferences may induce employers to *patronize* the favored workers, which in turn may undercut their incentives to acquire necessary skills.

We have shown that a policy of subsidizing workers directly for achieving employment success can generally achieve the elimination of prejudicial views about minorities without the negative side-effects possible under affirmative action. This result has an important practical implication: if one objective in the fight against discrimination is to break down stereotypes, then it will sometimes be better to encourage disadvantaged workers to supply greater effort, than to bribe or coerce employers into promoting these workers.

REFERENCES

Aigner, Dennis J. and Cain, Glen G., "Statistical Theories of Discrimination in the Labor Market," *Industrial and Labor Relations Review*, January 1977, *30*, 175–87.

Akerlof, George, "The Economics of Caste and of the Rat Race and Other Woeful Tales," *Quarterly Journal of Economics*, November 1976, *90*, 599–617.

Arrow, Kenneth J., "The Theory of Discrimination," in Orley Ashenfelter and Albert Rees, eds., *Discrimination in Labor Markets*, Princeton, NJ: Princeton University Press, 1973, pp. 3–33.

Becker, Gary S., *The Economics of Discrimination*, Chicago: University of Chicago Press, 1957.

Borjas, George J. and Goldberg, Matthew S., "Biased Screening and Discrimination in the Labor Market," *American Economic Review*, December 1979, *68*, 918–22.

Coate, Stephen and Tennyson, Sharon, "Labor Market Discrimination, Imperfect Information and Self Employment," *Oxford Economic Papers*, April 1992, *44*, 272–88.

Glazer, Nathan, *Affirmative Discrimination: Ethnic Inequality and Public Policy*, New York: Basic Books, 1975.

Kahn, Lawrence M., "Customer Discrimination and Affirmative Action," *Economic Inquiry*, July 1991, *26*, 555–71.

Lang, Kevin, "A Language Theory of Discrimination," *Quarterly Journal of Economics*, May 1986, *101*, 363–82.

_____, "A Sorting Model of Statistical Discrimination," mimeo, Boston University, 1990.

Leonard, Jonathan S., "The Impact of Affirmative Action on Employment," *Journal of Labor Economics*, October 1984, *2*, 439–63.

Loury, Glenn C., "Why Should We Care About Group Inequality?" *Social Philosophy and Policy*, Autumn 1987, *5*, 249–71.

Lundberg, Shelly J., "The Enforcement of Equal Opportunity Laws Under Imperfect Information: Affirmative Action and Alternatives," *Quarterly Journal of Economics*, February 1991, *106*, 309–26.

_____ **and Startz, Richard,** "Private Discrimination and Social Intervention in Competitive Labor Markets," *American Economic Review*, June 1983, *73*, 340–7.

Milgrom, Paul and Oster, Sharon, "Job Discrimination, Market Forces, and the Invisibility Hypothesis," *Quarterly Journal of Economics*, August 1987, *102*, 453–76.

Phelps, Edmund S., "The Statistical Theory of Racism and Sexism," *American Economic Review*, September 1972, *62*, 659–61.

Schotter, Andrew and Weigelt, Keith, "Asymmetric Tournaments, Equal Opportunity Laws, and Affirmative Action: Some Experimental Results," *Quarterly Journal of Economics*, May 1992, *107*, 511–39.

Smith, James P. and Welch, Finis, "Affirmative Action and Labor Markets," *Journal of Labor Economics*, April 1984, *2*, 269–301.

Spence, Michael A., *Market Signaling: Information Transfer in Hiring and Related Screening Processes*, Cambridge, MA:

Harvard University Press, 1974.

Starrett, David, "Social Institutions, Imperfect Information, and the Distribution of Income," *Quarterly Journal of Economics*, May 1976, *90*, 261–84.

Weiss, Andrew, "Determinants of Quit Behavior," *Journal of Labor Economics*, July 1984, *2*, 371–87.

Welch, Finis, "Employment Quotas for Minorities," *Journal of Political Economy*, August 1976, *84*, S105–39.

_____, "Affirmative Action and Discrimination," in Steven Shulman and William Darity, Jr., eds., *The Question of Discrimination*, Middletown, CT: Wesleyan University Press, 1989, pp. 153–89.