Must positive discrimination be permanent? The impact of implementing and removing affirmative-action programs on workers’ skill acquisition

Nick Feltovich∗
Department of Economics
Monash University
Clayton VIC 3800, Australia
nicholas.feltovich@monash.edu

Lata Gangadharan
Department of Economics
Monash University, Clayton Campus
VIC 3800, Australia
Lata.Gangadharan@buseco.monash.edu.au

Michael P. Kidd
School of Accounting, Economics and Finance
Deakin University, Melbourne Campus
VIC 3125, Australia
mkidd@deakin.edu.au

June 21, 2011

Abstract

The impact of affirmative action on the behavior of targeted individuals is theoretically indeterminate: it can either raise or lower their investment in skill acquisition, thus diminishing or exacerbating a pre-existing negative stereotype. The impact of removing an existing affirmative-action program is similarly indeterminate. This paper uses a laboratory experiment to examine skill acquisition before and after affirmative action is introduced, and after its subsequent removal. We consider three parameterizations of a statistical discrimination model, with differing theoretical implications. Our results show qualitative treatment effects that are mostly consistent with theoretical predictions.

Journal of Economic Literature classifications: J15, J78, J71, D81.

Keywords: statistical discrimination, signaling, positive discrimination, affirmative action, experiment

*Corresponding author. Funding for this research was provided by Deakin University’s Faculty of Business and Law and by University of Melbourne. Some of this research took place while Feltovich was at University of Aberdeen. We thank Jay Pot for his assistance in programming the experiment, and Paul Carlin, Robert Moffitt, Ron Oaxaca and Jörg Oechssler for helpful suggestions and comments.
1 Introduction

The term affirmative action is used to describe policies aimed at improving the welfare of particular target groups in the population, generally on the basis of ethnic or racial status, sex, religious affiliation or caste. The focus is typically on provisions to address underrepresentation in employment, access to education, or political power, and it is often rationalized as a way of correcting prior discrimination or disadvantage. Affirmative action is both topical and controversial. Most developed countries, and many developing countries, have affirmative-action policies in place, including the United States, which has had them at federal level since the early 1960s. However, affirmative-action policies have come under attack, particularly in the U.S., in a string of recent state-level legislation and high-profile court cases. California’s Proposition 209, passed by referendum in 1996, banned all forms of affirmative action at state level. The states of Washington and Florida followed suit in 1998 and 2000 respectively. Recent U.S. Supreme Court cases such as Gratz v. Bollinger and Grutter v. Bollinger (both in 2003) tested the constitutionality of using race or minority status as a legitimate criterion in university admissions. While the Supreme Court upheld the University of Michigan Law School’s admissions policy—which adopts race as one aspect of a multi-dimensional set of criteria—it did so on the grounds that race was used in the interest of maintaining diversity (following the precedent of the Regents of the University of California v. Bakke (1978) decision). By contrast, simple quotas and blanket points awards in the admissions process based on race (such as the one used by University of Michigan for undergraduate admissions) were judged to be unconstitutional.

Given the likelihood that pressure against color–sighted affirmative–action policies will continue to build in at least some countries and regions, one must acknowledge the possibility that policy makers will respond by dismantling existing affirmative–action programs. With that possibility in mind, it becomes increasingly important to understand what the effects of such a change would be. Note that the effects of repealing an existing affirmative–action policy are not necessarily the same as the effects of never having had such a policy, since it is likely that employers’ and institutions’ beliefs about workers’ attributes adjust over time, and thus depend not only on policies in effect currently, but also past policies. For the same reason, the effects of repealing an affirmative–action policy cannot be assumed to be the diametric opposite of the effects of enacting such a policy.

To understand the effects of removing affirmative action, it is important to understand why discrimination occurs in the first place. There are several reasons why discrimination might arise and persist in the absence of affirmative action (Becker, 1972; Phelps, 1972; Lundberg and Startz, 1983; Coate and Loury, 1993; Fryer and Loury, 2005). Statistical discrimination (Phelps, 1972; Arrow, 1973) provides one promising explanation. Statistical theories of discrimination focus on the idea that employers may rationally use workers’ observable characteristics—such as ethnic group or sex—as a proxy for the workers’ underlying

---

1 As a developing–country example, affirmative–action policies in India have existed in some form for over a century. Presently, almost a quarter of government jobs and places in state–funded colleges are reserved for members of the lowest caste, and a bill introduced in December 2006 extended affirmative action to include other low–caste groups and increased the “quota” to almost 50 percent. Prakash (2009) estimates the effect of these programs on labor–market and other outcomes. Also, Malaysia has an affirmative action program providing preferential access to education to the native Malays; Fang and Norman (2006) examine this program’s effects on labor–market outcomes. Note that affirmative action can also be implemented outside of labor markets; for example, Marion (2007, 2009) and Krasnokutskaya and Seim (2011) study affirmative action in government procurement.
ability or competence to perform a task, both of which are typically ex ante unobservable. In that case, initial employer perceptions of differences in ability between groups (whether justified or not) can become self-fulfilling, through their effects on workers’ incremental returns to additional training or education. Specifically, an individual belonging to a group with a lower initial assessment of ability might face a lower return to skill acquisition, and thus little incentive to invest. This in turn reinforces the employer’s original assessment and hence the negative stereotype.

Empirical exploration of statistical discrimination, however, is inherently difficult as key variables such as firms’ beliefs about negative stereotypes (the belief that one group is less productive than others) are not directly observable, and difficult even to infer. Additionally, natural experiments involving affirmative-action programs are rare—and those involving the removal of such programs are extremely rare—making it difficult to examine such changes directly.\(^2\) Laboratory experiments with human subjects, on the other hand, allow for control over many aspects of the environment in which firms and workers make decisions. In particular, it is easy in the lab to simulate the impact of changes in affirmative-action policy, while keeping everything else the same, in order to isolate and clarify the impact of these policies.

In this paper, we use a laboratory experiment to examine the effects of affirmative action. Specifically, we focus on its impact on the behavior of individuals in the target group, who initially face a negative stereotype. We examine the pattern of worker skill acquisition upon the introduction of an affirmative-action policy and after its subsequent removal. The experiment is designed to test three central ideas relating to affirmative action. These ideas parallel theoretical implications arising from the stylized model in Coate and Loury’s (1993) seminal paper, where firms choose whether to allocate a newly-hired worker to a high-skill, high-wage task or to a low-skill, low-wage task. First, consider a prevailing negative stereotype, with firms believing that the proportion of members of one group who are qualified for the high-skill, high-wage, task is lower than that of the other group. Coate and Loury demonstrate that, depending upon the values of parameters that characterize this stereotype, the introduction of affirmative action can lead to either a “benign” or “patronizing” equilibrium. In the benign equilibrium, the negative stereotype disappears, and the proportions of advantaged and disadvantaged workers who are qualified for and allocated to the higher-level task are equalized. In the patronizing equilibrium, the proportion of the disadvantaged group that is qualified for and allocated to the high-wage task remains lower than that of the advantaged group; that is, the negative stereotype continues to exist. Thus the theoretical impact of introducing affirmative action is ambiguous; this ambiguity is of obvious policy relevance.

The second key issue is that for some parameter combinations yielding a patronizing equilibrium, affirmative action is predicted to lead to a worsening of the negative stereotype, with an even smaller proportion of the disadvantaged group investing in skills than before affirmative action was introduced. The intuition underlying this theoretical result is the widely-cited potential problem with affirmative action: the disincentive effects arising from favorable treatment. If the target group knows that they must be employed according to some legislated quota, this might lead to a disincentive for those in that group to exert effort. In the theoretical model, this is captured by a disincentive to invest in skills and become qualified for the

\(^2\)Myers (2007) treats the introduction of California’s Proposition 209 as a natural experiment. Her results suggest that its enactment—that is, the removal of an affirmative action program—led to a sharp drop in minority employment in California relative to the rest of the US.
higher–level task. This possibility ties neatly into the broader context of Heilman et al. (1987), in which individuals in the newly–favored group lose confidence and lower perceptions of their own performance.

The third key issue is that, depending upon parameters, a temporary affirmative–action intervention may or may not have a permanent impact. From an initial negative–stereotyping equilibrium, the introduction of affirmative action leads theoretically to a new equilibrium. In the case of a benign equilibrium, the gains from affirmative action remain even after dismantling of the affirmative–action program: the negative stereotype has disappeared, so that both groups are treated symmetrically by firms and thus behave in similar ways. For many patronizing equilibria, however, removing the affirmative–action program results in those gains being lost: the negative stereotype continues to exist and to drive firm decisions, and hence worker investment behavior. Once again, therefore, the policy implications of affirmative action are theoretically unclear: in order to maintain equal proportional representation of advantaged and disadvantaged groups in the higher–level task, it might be that affirmative action needs to be permanent, or alternatively it might need only be a temporary intervention, depending on parameters.

In our experiment, subjects play the role of employees from the disadvantaged group, and the employers’ decisions (along with those of employees from the advantaged group) are automated. We compare investment behavior under three policy treatments. In the first treatment, subjects make investment decisions in a pre–affirmative–action environment, under an existing negative stereotype and with no affirmative–action program in place. In the second treatment, they make investment choices under an affirmative–action program—with firms required to allocate equal proportions of the two groups to the high–wage task—and in the third treatment, the program is removed. We also consider three parameter combinations that vary qualitatively in their predicted effects of these policies. In a “benign treatment”, parameters are chosen to give rise to a benign equilibrium, with affirmative action predicted to lead to increased investment in skills by the disadvantaged group, up to parity with the advantaged group, hence eradicating the negative stereotype held by employers. These effects persist even if the affirmative–action policy is subsequently repealed, so that in this case, a temporary intervention has a permanent effect. Our other two treatments involve parameters chosen to lead to a patronizing equilibrium, where affirmative action does not eliminate the negative stereotype. In one of these, which we call our “patronizing treatment”, affirmative action is predicted to lead to investment by the disadvantaged group increasing, but remaining below the level of the advantaged group, thus perpetuating the stereotype. In the other, our “worsening treatment”, the disadvantaged group actually becomes less likely to invest as a result of affirmative action, exacerbating the negative stereotype. In either of these latter treatments, repealing affirmative action undoes its effects, meaning that it would need to be a permanent feature in order to have a lasting impact on workers’ labor–market outcomes.

Experimental research that investigates job market signaling and statistical discrimination is not new; in fact, there is now a small literature devoted to these topics (see, for example, Feltovich and Papageorgiou, 2004; Fryer et al., 2005; Healy, 2007; Kübler et al., 2008). There has even been a previous experimental

3Note the distinction we make here between “patronizing equilibrium” and “patronizing treatment”. Our experimental design comprises both a patronizing treatment and a worsening treatment—both of which have patronizing equilibrium as a theoretical implication—along with a benign treatment in which the benign equilibrium is predicted by the theory.

4A related strand of literature uses experiments to look at the effects of affirmative–action programs on effort exerted by both favored and unfavored worker types in tournaments (Schotter and Weigelt, 1992); women’s entry into tournaments (Niederle, Segal
study of affirmative action in a statistical–discrimination context (Kidd et al., 2008). The contribution of our paper to this literature is twofold. First and most importantly, we attempt to determine the effects of not only introducing affirmative action (as Kidd et al. also did), but also subsequently dismantling the program; to our knowledge, no other experimental study has attempted to do this. Our examination of both the introduction and subsequent abolition of affirmative action addresses a key policy issue: whether affirmative–action policies would need to continue indefinitely or whether exposure to such policies for a limited period of time can help in eliminating discrimination in labor markets. Second, we assess the robustness of Kidd et al.’s conclusions regarding the initial imposition of affirmative action, by examining an important case that they did not consider: namely, our worsening treatment, in which negative stereotypes are predicted to become worse as a result of affirmative action. Ignoring this possibility risks reaching overly optimistic policy conclusions about the beneficial effects of affirmative action, and conversely, overly negative conclusions about the harmful effects of removing it.

The results from our experiment give mixed support to the theory. Where our focus is on theoretically–derived point predictions (e.g., when the theory predicts a certain average level of investment by disadvantaged workers), the data are typically inconsistent with the theory. In particular, there is substantial overinvestment by individuals from the disadvantaged group throughout the experiment (see Section 5 for some conjectures about the cause of this result). On the other hand, the directional predictions arising from the theory (e.g., a prediction of an increase in investment when affirmative action is implemented) are largely borne out in the data. The implication of our results is that whether there is a need for a permanent affirmative–action policy depends to some extent on the nature of the original negative stereotype held by firms, but any such need might be less than the theory would imply.

2 Theoretical framework

Our testable hypotheses come from Coate and Loury’s (1993) theoretical model of statistical discrimination and affirmative action. While this model, like most theoretical models, makes many simplifying assumptions, we believe that it captures many of the important aspects of discrimination in labor markets, and thus provides a useful starting point for our experimental analysis. In this section, we summarize the essential components of the model; the interested reader can refer to Coate and Loury (1993) for additional details. Consider a labor market with a large number of identical risk–neutral firms, and risk–neutral workers who each belong to one of two types, W and B; \( \lambda \) is the proportion of W types in the market. A representative firm is randomly matched with a set of new hires (drawn uniformly from the worker population), and then assigns each new hire to one of two tasks: Task 0 (\( T_0 \)) or Task 1 (\( T_1 \)). Only workers who have invested in the requisite skills are qualified for the higher–level task \( T_1 \), but all workers are qualified for \( T_0 \). Whether a particular worker is qualified for \( T_1 \) is that worker’s private information, and cannot be directly observed by firms. However, firms perfectly and costlessly observe the worker’s type, along with a test result that is correlated with the worker’s investment choice, prior to making its decision. This decision—the allocation

and Vesterlund, 2009; Balafoutas and Sutter, 2010) and the incentive effect of real–effort experiments on affirmative action (Cal- samiglia et al., 2009). See Anderson et al. (2006) for a survey of economics and psychology experiments involving discrimination in various forms.
of workers to tasks—is the only choice the firm makes: it does not choose which people to hire (or fire), nor does it set wages.

Each worker makes a binary decision of whether to undertake a costly investment in skills, which ensures qualification for, but not assignment to, \( T_1 \). All workers then take the test, which is costless, and each obtains a test score \( \theta \in [0, 1] \), with higher values more likely if the worker is qualified. The employer realizes a positive \( x_q \) (negative \( -x_u \)) return from assigning a qualified (unqualified) worker to \( T_1 \), net of the return from alternatively assigning the worker to \( T_0 \). Define \( r = \frac{x_q}{x_u} \) to be the ratio of net gain to net loss for the firm. Finally, assume that workers assigned to \( T_0 \) and \( T_1 \) receive payoffs (gross of any cost of investment incurred) of 0 and 1 respectively.

A given worker’s skill–acquisition choice depends on that worker’s cost of investment (which the worker can observe before making the investment choice, but which cannot be observed by firms), and on the probability of an investor versus non–investor being assigned to \( T_1 \), which depends in turn on the decision rule adopted by the employer—i.e., the employer standards that workers face. The employer starts with a prior belief \( \pi_i \) \((i = b, w) \) about the probability of a worker in the given group (B or W) being qualified. After observing the worker’s test score \( \theta \), the employer forms a posterior probability \( \xi \) which depends on \( \pi_i \) and the likelihood ratio \( \varphi(\theta) \) (the odds that a worker with a given score \( \theta \) is unqualified). If \( \xi \) is above a cutoff level (determined by the values of \( x_q \) and \( x_u \)), the worker will be assigned to \( T_1 \).

In the special case of Coate and Loury’s model that forms the basis of our experimental design (see Coate and Loury, 1993, pp. 1230–1232), a number of simplifying assumptions are made. There are only three possible test scores: pass, uncertain, or fail. (That is, the support of \( \theta \) is simply \( \{\theta_p, \theta_u, \theta_f\} \) with \( \theta_p > \theta_u > \theta_f \).) A worker who invests can earn either a pass or an uncertain (but not a fail), while a worker who chooses not to invest can earn either an uncertain or a fail (but not a pass). Thus, in the absence of external constraints such as affirmative action, the employer will choose to assign all workers with a pass to \( T_1 \), and all workers with a fail to \( T_0 \). For a worker with an uncertain test score, the task assignment depends upon the employer’s prior beliefs \( \pi_i \); the employer will assign a worker from group \( i \) with an uncertain score to \( T_1 \) if \( \pi_i > \frac{\varphi}{\varphi + 1} \equiv \hat{\pi} \), and to \( T_0 \) if \( \pi_i < \hat{\pi} \). As \( r \) depends on the exogenously set \( x_q \) and \( x_u \), and \( \varphi \) depends on the exogenously set probabilities of receiving pass, uncertain, or fail scores conditional on investing or not investing, \( \hat{\pi} \) is exogenously fixed. So generically, the employer optimally chooses either to assign all members of a group with uncertain test scores (a liberal standard) to \( T_1 \) or to assign none of them (a conservative standard) to \( T_1 \).

For their part, workers invest if their expected payoff net of costs is greater than zero. When workers expect the liberal (conservative) employer standard for their group, enough (so few) invest that the employer’s standard is optimal. As in the general case, a (pre–affirmative action) negative–stereotyping equilibrium is a pair of employer beliefs \( (\pi_w, \pi_b) \) that are confirmed by the proportions of W and B workers investing; that is, with \( \pi_w > \hat{\pi} > \pi_b \).

Under affirmative action, firms are required to assign at least as large a proportion of B workers to \( T_1 \)

---

5As Coate and Loury (1993) show, there is a close relationship in this version of the model between equilibrium firm beliefs and the likelihood of passing the test. In a negative–stereotyping equilibrium, it has to be the case that \( \pi_w \) is equal to the probability of receiving a fail score conditional on not investing, and \( \pi_b \) is equal to the probability of receiving a pass score conditional on investing. Thus, in equilibrium, firms’ stereotypes are not primitive, but rather are driven by the technology of testing.
as W workers. As a result, in the affirmative–action equilibrium, the employer optimally assigns all Bs with passing and uncertain scores to $T_1$, as well as a fraction $\alpha = \frac{\pi_w - \pi_b}{1 - \pi_b}$ of those with failing scores. In the special case we consider, Coate and Loury demonstrate that there are two classes of equilibria, depending on the original negative stereotype held by firms; denote this by $\pi_0^w$ and $\pi_0^b$. If $\pi_0^w > \pi_0^b$, $\pi_0^w > 0.5$, and $\lambda$ (the proportion of W workers) is sufficiently large, then there is a unique stable equilibrium (called the “patronizing” equilibrium) in which employers continue to possess negative stereotypes about B workers, which are correct in equilibrium, with $\pi_b = 1 - \pi_w < \pi_w$. Alternatively, if $\pi_0^b < \pi_0^w < 0.5$, there is a locally stable equilibrium (called the “benign” equilibrium) where negative employer stereotypes are gradually eliminated, so that eventually $\pi_w = \pi_b$.

3 The experiment

Following Feltovich and Papageorgiou (2004) and Kidd et al. (2008), we have designed a one–sided experiment, in which subjects make decisions in the role of workers, and the firms are computerized. This has two advantages: (1) we can “impose” appropriate firm beliefs rather than expecting human subjects in the role of firms to learn them, and (2) the results are expected to have less noise, as we can abstract away from issues such as social preferences between firms and workers, and strategic uncertainty (on the part of the workers) about firms’ actions. Additionally, W–type workers in the experiment are assumed to behave as in the Coate–Loury special case described above; i.e., they invest at the rate that precisely matches and thus confirms firm beliefs. This allows us to focus on the B–type workers and their investment behavior.

A round of the experiment begins with B–type workers facing a decision of whether or not to invest. The parameters determining the probabilities of the test scores (pass, uncertain, and fail) conditional on investment choice are determined by the initial parameter selection $(\pi_0^w, \pi_0^b)$. After the worker’s investment decision is made, the computer indicates the realization of the test score (Pass, Uncertain, or Fail), as well as the task assignment (Task 1 or Task 0). Finally, the computer updates firm beliefs regarding the proportion of workers who are qualified.

This last aspect of the round (updating of beliefs) requires further discussion. The focus of the Coate–Loury model is on the impact of introducing affirmative action on an established negative stereotyping equilibrium (with $\pi_0^w > \pi_0^b$). We use the pre–AA rounds of the experiment to create an environment in which a negative stereotype exists, in the expectation that by the end of these rounds, B workers’ decisions reflect this. To create the negative stereotype, the optimal hiring standard is assumed fixed and a function of $\pi_0^w$ and $\pi_0^b$. Since the probabilities of earning particular test scores conditional on the investment decision are also fixed, this implies that the scenario encountered by workers in each round of pre–AA is stationary and independent of aggregate investment behavior. Nonetheless, interpreting $(\pi_0^w, \pi_0^b)$ as an equilibrium within the experimental context is problematic, since we are not guaranteed that experimental subjects in the role of workers will invest in a manner consistent with the model. Behavior may deviate from the theoretical prediction for several reasons, including optimization errors and experimenter–induced demand effects.

In the first round in which affirmative action is in place, the parameters $(\pi_w, \pi_b)$ are assumed to be equal to the initial parameters $(\pi_0^w, \pi_0^b)$ from the pre–AA treatment. In later AA rounds, firms update their beliefs as follows: $\pi_b^{t+1}$ is equal to the proportion of Bs investing in round $t$, subject to a cap of $\pi_w^t$, while
\( \pi_w^t = \pi_b^0 \) in all rounds. This in turn implies that under affirmative action, the proportion of the B group that is patronized (i.e., the proportion of Bs failing the test but nevertheless assigned to task \( T_1 \)) in round \( t \) is

\[
\alpha = \frac{\pi_w^t - \pi_b^0}{\pi_b^0},
\]

which of course will vary with \( \pi_b^0 \).

An important innovation of our experiment is the post–AA treatment, which was designed to help us understand the permanency requirement of affirmative–action intervention. As noted in Section 2, Coate and Loury demonstrate that the need for permanent affirmative action depends only on model parameter values. Specifically, for parameters yielding a patronizing equilibrium, if the theoretical prediction for \( \pi_b \) after the imposition of affirmative action (i.e., \( 1 - \pi_b^0 \)) is less than \( \hat{\pi} = \frac{\delta}{\gamma + \phi} \), then in order to maintain equal proportional representation of the two groups, affirmative action must be continued indefinitely.

### 3.1 Experimental design and procedures

Our experiment utilizes a 3x3 design. The first treatment variable comprises the model parameter pairs we use: \( (\pi_w^0, \pi_b^0) = (.4, .1), (.8, .1), \) and \( (.8, .4) \). Note that in all three, the condition \( \pi_w^0 > \pi_b^0 \) holds, and \( r = \frac{\pi_w}{\pi_b} \) is assumed to be chosen so that \( \pi_w^0 > \hat{\pi} > \pi_b^0 \). We will refer to these parameter pairs as our *model treatments*, to distinguish them from our *policy treatments* of pre–AA (before affirmative action is enacted), AA (while affirmative action is in place), and post–AA (after affirmative action is repealed). Also, we will use the term “cell” to refer to a combination of model treatment and policy treatment (e.g., the cell with \( (\pi_w^0, \pi_b^0) = (.4, .1) \) and post–AA).

The first of our model parameter pairs gives rise to a benign equilibrium: we have \( \pi_w^0 = 0.4 < 0.5 \), so that upon introduction of affirmative action, \( \pi_b \) is predicted to rise to \( \pi_b^0 = 0.4 \). The temporary intervention of imposing affirmative action in this “benign treatment” should therefore have a permanent impact; that is, even after the repeal of affirmative action, equality between the two groups will be maintained. The second model treatment gives rise to a patronizing equilibrium. We have \( \pi_w^0 = 0.8 > 0.5 \) and \( \pi_b^0 = 0.1 < 0.2 = 1 - \pi_w^0 \), so that after affirmative action is imposed in this “patronizing treatment”, \( \pi_b \) should rise to \( 1 - \pi_w^0 = 0.2 \), which is assumed to be less than \( \hat{\pi} \). This means that affirmative action will theoretically need to be a permanent feature in order to have a lasting effect. The third model treatment also leads to a patronizing equilibrium. Here, \( \pi_w^0 = 0.8 > 0.5 \) and \( \pi_b^0 = 0.4 > 0.2 = 1 - \pi_w^0 \), so that \( \pi_b \) is projected to fall to \( 1 - \pi_w^0 = 0.2 \) which is also less than \( \hat{\pi} \); that is, investment by B workers actually falls as a result of affirmative action. In this “worsening treatment”, affirmative action will again need to be permanent in order to have a lasting effect. Table 1 summarizes the parameter combinations and theoretical predictions for each of our treatments. In this table, it is important to note that the predicted frequencies displayed in the

<table>
<thead>
<tr>
<th>Model treatment</th>
<th>( \pi_w^0 )</th>
<th>( \pi_b^0 )</th>
<th>Predicted B invest. freq.</th>
<th>Predicted change, B invest. freq.</th>
<th>Sessions/subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign</td>
<td>0.4</td>
<td>0.1</td>
<td>0.1</td>
<td>0.4</td>
<td>Increase</td>
</tr>
<tr>
<td>Patronizing</td>
<td>0.8</td>
<td>0.1</td>
<td>0.1</td>
<td>0.4</td>
<td>Increase</td>
</tr>
<tr>
<td>Worsening</td>
<td>0.8</td>
<td>0.4</td>
<td>0.2</td>
<td>0.1</td>
<td>Decrease</td>
</tr>
</tbody>
</table>
table do not represent mixed strategies played by the B workers. Rather, they imply corresponding threshold investment costs, with a given worker choosing to invest if and only if his realized investment cost is less than the threshold (chosen so that the probability of the actual cost being below the threshold is equal to the predicted investment frequency shown).

As shown in the table, there were a total of nine experimental sessions, three with each set of model parameters. Subjects in the sessions were recruited from the undergraduate and graduate student bodies at Deakin University in Melbourne, with the majority of participants drawn from the Faculty of Business and Law. No subject participated in more than one session.

At the beginning of a session, subjects were seated at individual computer terminals so that they could not observe others’ computer screens. Each session began with subjects using pen and paper to complete a background history and answer a few lottery–choice questions in order to provide information about their risk attitudes. Subjects then participated in the main, computer–based portion of the experiment. Their computer screens displayed the experimental instructions, which were also read aloud by the experimenter. Each session involved ten subjects, who played under one set of model parameters. That is, the model treatment (benign, patronizing, worsening) was varied between subjects, whereas the policy treatment (pre–AA, AA, post–AA) was varied within–subject. Each experimental session comprised a total of 37 rounds, in the following order: 1 practice pre–AA round, 10 paying pre–AA rounds, 1 practice AA round, 15 paying AA rounds, 10 paying post–AA rounds. (Subjects were informed at the beginning of a practice round that it would not count for payment.) Subjects were not told how many rounds would make up each portion of the session, nor were they told the total number of rounds; instead, they were merely told that the maximum time for the entire session was 90 minutes. Sessions lasted between 75 and 90 minutes in total, with at least 60 minutes devoted to the main portion of the experiment.

We further divided each round of a session into ten “periods”. At the beginning of a new round, subjects are presented with a binary choice—to invest or not. The subject’s decision is then binding for all ten periods within the round. Subjects are informed that firms have imperfect information (specifically, they could not observe the investment decision), and hence the computer would select a test result in each period (i.e., ten separate tests per round), to determine which task(s) the subject will be assigned to. A subject’s investment cost was constant over all ten periods within a round, but the ten test results were chosen i.i.d., and the resulting allocation of the worker by the firm to Task 0 or Task 1 was chosen independently across the ten periods. The purpose of this division of rounds into periods was our expectation that making subjects observe results from ten tests for each investment choice would improve their understanding of the environment and thus reduce decision errors.

The payoff in each period, gross of any investment cost incurred, for subjects assigned to Task 1 was set to 150 points, and the payoff for Task 0 was set to 50 points. Subjects were informed at the beginning of a round of the probability of being assigned to either task given their investment decision—that is, they did not have to forecast the firm’s hiring standard—and were notified when the policy treatment changed (from pre–AA to AA, or from AA to post–AA) of any changes to the hiring standard. Subjects were also informed at the beginning of a round of their cost of investment in that round (drawn from a uniform distribution over \{0, 1, 2, ..., 100\}, i.i.d. for each subject and round).6 Once the investment choice was made in a given round,
the subject received feedback including the assignment (Task 0 or Task 1) and the payoff in points for each period. Throughout the experiment, subjects had access to the history of their results from previous rounds. Subjects were not given information about the choices or payoffs of other subjects.

A subject’s total payment for the experiment was determined as follows: the computer randomly selected five of the paying (non–practice) rounds, then the net payments for these rounds were summed and divided by 1000 to convert to Australian dollars (AUD). This sum, which could in theory be positive or negative (but which in practice was always positive), was added to a participation fee of AUD 20 and rounded to the nearest 10 cents. Overall, subjects’ earnings ranged from AUD 27.50 to AUD 42.50. Experimental materials (including items such as general instructions and the set of questions about background and risk profile) are available from the authors upon request.

In an attempt to reduce demand effects arising from potentially emotionally–loaded terms such as “discrimination” and “affirmative action”, we tried to make the experiment relatively context–free. Subjects were not provided with any information about the negative stereotyping environment or the introduction of affirmative action; rather, the instructions described the situation as an “investment game”. From the subjects’ perspective, the switch from the pre–AA to the AA treatment, and thence to the post–AA treatment, was reflected only in a change in the firms’ hiring standard and associated probability of being assigned to Task 1 contingent on test performance and investment decision.

3.2 Hypotheses

Using our experimental design, we aim to examine the following theoretical propositions, all of which are derived from Coate and Loury’s (1993) analysis and have been described above:

**Hypothesis 1** In the benign treatment, introducing affirmative action leads to an increase in B worker investment (relative to pre–AA), while removing it has no effect (relative to AA).

**Hypothesis 2** In the patronizing treatment, introducing affirmative action leads to an increase in B worker investment (relative to pre–AA), while removing it leads to a decrease (relative to AA).

**Hypothesis 3** In the worsening treatment, introducing affirmative action leads to a decrease in B worker investment (relative to pre–AA), while removing it leads to an increase (relative to AA).

4 Experimental results

In the first subsection below, we begin with some discussion of treatment–wide aggregates, where we assess significance based on nonparametric tests requiring minimal distributional assumptions and based on only one independent summary statistic value per session. In the second subsection, we present results that are partly disaggregated, either by the round number (thus showing dynamics) or according to other variables.

---

7See Siegel and Castellan (1988) for descriptions of the nonparametric tests used in this paper, and see Feltovich (2005) for critical values for the robust rank–order test used below.
In the third subsection, we report estimates from multivariate parametric regression models, in order to disentangle the contributions of our main treatment variables on investment behavior from those of various other variables. Finally, a fourth subsection collects the main results of our experiment and, when applicable, compares them to the theoretical hypotheses presented above.

4.1 Aggregate results

Aggregate investment frequencies over all non–practice rounds are shown in Table 2 for each cell—that is, each combination of model treatment and policy treatment—of the experiment (recall that rounds 1 and 12 were practice rounds for pre–AA and AA respectively). Also shown are the investment frequencies from the last three rounds of each policy treatment, which we use as a measure of behavior after subjects have gained some experience in a situation, as well as the corresponding theoretical predictions from the Coate–Loury model.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Frequency of investment (in %)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All rounds</td>
<td>Last 3 rounds</td>
<td>Equil.</td>
<td>All rounds</td>
<td>Last 3 rounds</td>
</tr>
<tr>
<td>Benign</td>
<td>45.0</td>
<td>43.3</td>
<td>10</td>
<td>50.9</td>
<td>43.3</td>
</tr>
<tr>
<td>Patronizing</td>
<td>41.0</td>
<td>31.1</td>
<td>10</td>
<td>35.6</td>
<td>26.7</td>
</tr>
<tr>
<td>Worsening</td>
<td>59.0</td>
<td>63.3</td>
<td>40</td>
<td>45.8</td>
<td>37.8</td>
</tr>
</tbody>
</table>

Two features of the experimental data stand out. First, there is substantial across–the–board overinvestment; in each cell, the observed frequency of investment is higher than the theoretical prediction, and the difference varies from 8–37 percentage points, with an experiment–wide average difference of almost 19 percentage points. If we pool the three model treatments, the evidence of overinvestment becomes stark: investment frequencies over all rounds are significantly higher than their predicted values in all three policy treatments (two–tailed Wilcoxon signed–rank test, pooled model treatments, all rounds, \( p < 0.01 \) for pre–AA and AA, \( p \approx 0.02 \) for post–AA). If we concentrate on the final three rounds of each policy treatment, the result is almost as strong: investment frequencies are significantly higher than those predicted in pre–AA (\( p < 0.01 \)) and post–AA (\( p \approx 0.04 \)), though the difference is insignificant in the AA treatment (\( p \approx 0.30 \)).

Second, within each policy treatment, differences across model treatments in observed frequencies tend to reflect, at least in a qualitative sense, the differences in theoretical predictions. In the pre–AA stage, the frequency of investment in the worsening treatment is significantly higher than in the other two model treatments (robust rank–order test, session–level data, pooled benign and patronizing treatments, \( \hat{U} = -2.91, p < 0.05 \)), matching the qualitative difference in their predicted frequencies of 0.4 versus 0.1. In the AA stage, the frequency of investment is higher in the benign treatment than in the other two model treatments, qualitatively consistent with the predictions of 0.4 versus 0.2, though this observed difference is statistically insignificant (pooled patronizing and worsening treatments, \( \hat{U} = -1.50, p > 0.10 \)). In the post–AA
stage, investment in the patronizing treatment is significantly less than in the other two model treatments (pooled benign and worsening treatments, $\bar{U} = +\infty$, $p \approx 0.012$), again reflecting the qualitative difference in predicted values of 0.1 and 0.4.

### 4.2 Disaggregated results and round–by–round dynamics

Figures 1 and 2 show how investment frequencies evolve over time in the three model treatments, along with the theoretical predictions for each cell. Also shown are the investment frequencies for the simulated W workers, which by construction are equal to their theoretically predicted values (note also that in the benign treatment, predicted investment frequencies are the same for both types of worker in the AA and post–AA treatments). In nearly all of the cells, observed investment frequencies begin well above their corresponding theoretical predictions, but decline on average over time, thus moving toward the predicted values. However, in most cases, observed frequencies remain above their predicted values, even in the last round of a policy treatment.

We can also see the effects on B workers’ investment choices arising from introducing affirmative action, and from subsequently removing it, in these figures. In all three model treatments, introducing affirmative action leads to an apparent change in investment that qualitatively matches the corresponding change in theoretical predictions: an increase in the benign and patronizing treatments, and a decrease in the worsening treatment. On the other hand, removing affirmative action leads to an apparent increase in investment in all three model treatments, despite the theoretical prediction being an increase in only the worsening treatment (recall that the theory implies no effect in the benign treatment and a decrease in the patronizing treatment).

The aggregate data above disguise a fine sensitivity of Bs’ choices to the realized cost of investment. Figure 3 gives a first impression of the relationship between investment cost and investment frequency; for each realization of the investment cost, this figure shows the fraction of times in the experiment in which a B worker with that cost chose to invest. It is only an illustration of this relationship, since the figure does not disaggregate by cell (and hence according to the predicted threshold cost at which the worker should switch
Figure 2: Frequency of B–type workers’ investment, patronizing and worsening treatments, all rounds (Note: Rounds 1 and 12 were practice rounds)

<table>
<thead>
<tr>
<th></th>
<th>Patronizing treatment</th>
<th>Worsening treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pre–AA</td>
<td>AA</td>
</tr>
<tr>
<td>Observed frequency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted frequency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Round</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3: Scatter–plot of investment cost and investment frequency, all cells, all non–practice rounds

from investing to not investing). Nonetheless, we can see a clear negative relationship between the cost of investment and the likelihood of investment.

This relationship is shown in more detail in Figure 4, which does plot investment cost and investment frequency separately for each cell of the experiment. To minimize noise due to small sample sizes, we re–aggregate the investment cost data partially, into intervals of costs: between 0 and 10, between 11 and 20, between 21 and 30, and so on. The negative relationship between cost of investment and frequency of investment seen in Figure 3 is apparent in all cells. However, we do not see stark differences across policy
treatments within any of the model treatments, in contrast to the theoretically–predicted differences that can be observed in Table 1. Also, we do not see compelling evidence that investment behavior in any of the cells can be characterized by a threshold cost—with workers in a given cell choosing to invest if and only if the realized investment cost is below the threshold—contrary to what is implied by the theory.

4.3 Parametric statistics

We next report the results of probit regressions, estimated in an effort to disentangle the effects of enacting and removing affirmative action in our three model treatments from the effects of other variables. For each model treatment (benign, patronizing, worsening), we estimate two probit models: a restricted model looking for average effects on investment, and an unrestricted model intended to detect time–varying effects. The dependent variable is the same in all models: an indicator variable equal to one if and only if the subject chooses to invest. The right–hand–side variables in the restricted model are a constant term, the realized cost of investment, and indicator variables for the pre–AA and post–AA treatments (so that the baseline is AA). The unrestricted model additionally uses the round number and its product with the pre–AA and post–AA indicators.8

All of the models were estimated using Stata (version 10), and incorporated individual–subject random effects. The results are shown in Table 3. Besides the usual estimated coefficients and standard errors for each variable, the table shows—for each of the unrestricted models—p–values corresponding to tests of joint significance for the two pre–AA variables and for the two post–AA variables.

The negative, and highly significant, coefficients for the investment cost in all six models are consistent with the negative relationship between this cost and investment frequency observable in Figures 3 and 4. Similarly, the negative coefficients for the round number suggest decreasing investment over time, as was

---

8We additionally estimated probits that also included subjects’ demographic characteristics and measures of risk attitude taken from the lottery–choice decisions from the beginning of the experimental session. The main results were not qualitatively affected by inclusion of these variables, so for space reasons we do not report the results here.
Table 3: Regression coefficients, standard errors, and significance results, non–practice rounds

<table>
<thead>
<tr>
<th>Variable</th>
<th>Benign</th>
<th>Patronizing</th>
<th>Worsening</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>1.633***</td>
<td>1.752***</td>
<td>1.328***</td>
</tr>
<tr>
<td></td>
<td>(0.173)</td>
<td>(0.375)</td>
<td>(0.220)</td>
</tr>
<tr>
<td>investment cost</td>
<td>–0.034***</td>
<td>–0.035***</td>
<td>–0.039***</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>pre–AA</td>
<td>–0.039</td>
<td>0.314</td>
<td>0.159</td>
</tr>
<tr>
<td></td>
<td>(0.116)</td>
<td>(0.409)</td>
<td>(0.125)</td>
</tr>
<tr>
<td>post–AA</td>
<td>0.013</td>
<td>2.129*</td>
<td>–0.028</td>
</tr>
<tr>
<td></td>
<td>(0.117)</td>
<td>(1.099)</td>
<td>(0.129)</td>
</tr>
<tr>
<td>round</td>
<td>–0.005</td>
<td>–0.039**</td>
<td>–0.043***</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td>(0.019)</td>
<td>(0.017)</td>
</tr>
<tr>
<td>pre–AA * round</td>
<td>–0.065</td>
<td>–0.048</td>
<td>–0.010</td>
</tr>
<tr>
<td></td>
<td>(0.036)</td>
<td>(0.039)</td>
<td>(0.036)</td>
</tr>
<tr>
<td>post–AA * round</td>
<td>–0.063</td>
<td>–0.014</td>
<td>–0.007</td>
</tr>
<tr>
<td></td>
<td>(0.036)</td>
<td>(0.040)</td>
<td>(0.036)</td>
</tr>
<tr>
<td>Joint test, pre–AA vars.</td>
<td>$p \approx .09$</td>
<td>$p \approx .06$</td>
<td>$p \approx .96$</td>
</tr>
<tr>
<td>Joint test, post–AA vars.</td>
<td>$p \approx .12$</td>
<td>$p \approx .12$</td>
<td>$p \approx .003$</td>
</tr>
<tr>
<td>pseudo–$R^2$</td>
<td>0.295</td>
<td>0.302</td>
<td>0.336</td>
</tr>
</tbody>
</table>

* (**, ***): Coefficient significantly different from zero at the 10% (5%, 1%) level.

seen in Figures 1 and 2; however, these coefficients are only significantly different from zero in two of the three model treatments.

The main result of these regressions is a somewhat surprising one: the coefficients for the pre–AA and post–AA indicators, and those for their products with the round number, are in most cases not significantly different from zero. At first glance, this suggests that subject behavior in the pre–AA, AA, and post–AA treatments is not distinguishable. However, some evidence that investment does vary across policy treatments comes from the joint tests, where the pre–AA variables are jointly significant at the 10% level in two of the three model treatments, while the post–AA variables are jointly significant at the 1% level in one model treatment and just miss being significant (at the 10% level) in the other two.

A better picture of the effects of imposing and then removing the affirmative–action program on behavior comes from estimates of the actual overall incremental effects of our pre–AA and post–AA treatments on investment, which are shown in Figure 5. In our unrestricted models, switching from the pre–AA treatment to the AA treatment affects the dependent variable through two distinct routes: through the change in the value of the “pre–AA” variable from one to zero, and through the change in the value of the “pre–AA * round” variable from the round number to zero. The combined effect of these changes on the value of the dependent variable is given by the expression

$$
\Phi (\bar{X} \cdot B) - \Phi \left( \bar{X} \cdot B + \beta_{\text{pre–AA}} + \beta_{\text{pre–AA} \times \text{Round}} \cdot t \right),
$$

where $\Phi$ is the normal c.d.f. used in the probit model, $\beta_z$ is the coefficient of the variable $z$, $\bar{X}$ is the row vector of the other right–hand–side variables’ values (depending on the variable, this will be taken to be
either the sample mean or the mean of an appropriate sub-sample), and $B$ is the column vector of their coefficients. A positive (negative) sign for this expression indicates an increase (decrease) in B workers’ investment due to implementing affirmative action. Similarly, the combined effect on investment due to switching from AA to post–AA (that is, from removing affirmative action) is given by

$$\Phi \left( \bar{X} \cdot B + \beta_{post-AA} + \beta_{post-AA+Round} \cdot t \right) - \Phi \left( \bar{X} \cdot B \right).$$

The estimated values for these expressions, along with 90% confidence intervals, in each round are shown in Figure 5. Also shown in each of the three panels is the horizontal segment corresponding to a zero effect.

Figure 5: Estimates of the effect of enacting/removing affirmative action on B workers’ investment probability (Circles represent point estimates; line segments represent 90% confidence intervals)

and dotted segments showing the theoretically predicted effects for the respective treatments (as shown in Table 1). In each panel of the figure, the first dark circle shows the initial effect of introducing affirmative action on B workers’ investment, compared to the hypothetical situation where it had not been introduced (so that pre–AA simply continued for another round). The trajectory of dark circles from round to round shows how this effect changes as time passes. Similarly, the first open circle shows the initial effect of repealing affirmative action on B workers’ investment—compared to the hypothetical case where affirmative action continued to be in place—and the trajectory of open circles shows how this effect changes over time.

Thus, in the benign treatment, the dark circles in Figure 5 are located substantially above the horizontal line segment representing a zero effect, and the confidence intervals are entirely above this zero line segment, implying a significant positive effect on B workers’ investment from introducing affirmative action in this treatment. The open circles, for their part, show a small, insignificant initial increase in investment due to removing affirmative action; over time, investment declines (though remaining insignificant). Both

---

9Note that we use 90% confidence intervals rather than the usual 95% confidence intervals in this figure. Because the hypothesized effects of introducing and removing affirmative action give rise—in all cases but one—to directional predictions, our rejection regions are one–tailed. Use of two–tailed 90% confidence intervals gives us rejection regions of 5% on each side.
effects are consistent with their corresponding theoretical predictions (an increase of 0.3 and no change, respectively).

In the patronizing treatment, enacting affirmative action also leads to an increase in B workers’ investment which is significantly greater than zero, but typically not significantly different from the theoretical prediction of +0.1. Eliminating affirmative action leads to an initial further significant increase in investment. Over time, this latter effect goes away, but the point estimate remains positive, and the effect remains significantly different from the theoretical prediction of −0.1.

In the worsening treatment, introducing affirmative action leads to an insignificant change in investment. While this change is a positive one, the confidence intervals are so wide that from round 5 on, it is statistically indistinguishable not only from zero, but also from the theoretical prediction of −0.2. As in the other treatments, removing affirmative action leads to an initial increase in investment that declines over time. This increase is significant in all rounds, though it is not significantly different from the theoretical prediction of +0.2.

### 4.4 Summary of results

In this section, we summarize the main results of the experiment. Many of these results parallel the hypotheses presented in Section 3.2. However, we have included a few additional noteworthy results.

**Result 1**  
*B workers’ overall investment frequencies are higher than their theoretically predicted values.*

This result, while independent of the hypotheses in Section 3.2, is obviously at odds with the theory, at least in terms of its point predictions. Support for this result can be found in the aggregate frequencies found in Table 2 and in the round–by–round frequencies seen in Figures 1 and 2, compared with the point predictions shown there and in Table 1.

**Result 2**  
*In all cells, B workers are less likely to invest as the cost of investment increases.*

This result is also not connected with any of the hypotheses in Section 3.2, though it is a weak implication of the theory. (The theoretical model makes a stronger prediction which is not observed in the data: that there is a cell–specific threshold value for the investment cost, such that the probability of investment should be one if the actual cost is lower, and zero if the actual cost is higher.) It is supported by the overall negative relationship between investment cost and investment frequency (seen in Figure 3), the similar negative relationships that were seen after disaggregating according to model and policy treatment (Figure 4), and the significant negative coefficients associated with the investment cost variable, seen in the regressions using the data from each of the three model treatments (Table 3).

**Result 3**  
*In the benign treatment, introducing affirmative action leads to a persistent increase in B worker investment, while removing it has no significant effect.*

This result, consistent with Hypothesis 1, is somewhat visible in the round–by–round descriptive statistics for the benign treatment (Figure 1), but its main support is found in the estimated effects of changing
policies presented in Figure 5 (left panel). Introducing affirmative action results in a significant initial increase in investment that continues to grow over time. Removing affirmative action, by contrast, leads to a positive but insignificant change initially, and while this incremental effect decreases over time until becoming negative, it remains insignificant.

**Result 4** *In the patronizing treatment, introducing affirmative action leads to a persistent increase in B worker investment, while removing it leads to a transitory increase.*

This result offers mixed evidence in favor of our Hypothesis 2, with the first half consistent with it and the second half inconsistent with it. The result is supported visually by the round–by–round descriptive statistics for the patronizing treatment (Figure 2, left panel), and statistically by the estimated effects seen in Figure 5 (middle panel). Investment increases following both the introduction and the removal of affirmative action, but only the former effect—consistent with the theory—remains significant over all rounds. (Recall that the theory predicted a *decrease* in investment following the repeal of affirmative action.)

**Result 5** *In the worsening treatment, introducing affirmative action has no effect on B worker investment, while removing it leads to an increase.*

This result offers mixed evidence for our Hypothesis 3, with the first half inconsistent with it and the second half consistent with it. The result can be seen in the round–by–round descriptive statistics for the worsening treatment (Figure 2, right panel), and is supported statistically by the estimated effects in Figure 5 (right panel). Introducing affirmative action leads to only a slight and insignificant increase in investment (in contrast to the theoretical prediction of a *decrease*), while removing it leads to a significant increase that, while becoming smaller over time, remains significant in all rounds—consistent with the theory.

## 5 Discussion

Recent legislation and legal challenges have called into question the long–term viability of color–sighted affirmative action policies—that is, explicit preferential treatment of a targeted racial group, intended to correct for past disadvantage. Also, theories of statistical discrimination suggest that under certain circumstances, affirmative action may lead to significant disincentive effects, such as a reduction in levels of investment in skills by the targeted group and an associated worsening of the original negative stereotype. Thus, there is reason to believe that the future of color–sighted affirmative action is in doubt, at least in some countries. As regional and national governments consider dismantling affirmative–action programs, it becomes increasingly important to understand the likely effects of such changes.\(^{10}\)

\(^{10}\)Our research has focused on the direct effects of eliminating an affirmative–action program, as opposed to examining alternative policies that are likely to take its place. It is worth noting that such alternative policies may well have their own drawbacks. For example, in 2003, the U.S. Department of Education advocated a race–neutral approach to university admissions, with an aim to improving racial diversity without using explicit racial preferences. However, recent research has expressed concern that this kind of color–blind policy may lead to an inefficient allocation of resources in higher education (Fryer et al., 2008; Ray and Sethi, 2009).
This paper presents the results from a novel experiment that enables a direct comparison of labor-market outcomes not only before and after implementation of an affirmative-action program, but also after a subsequent repeal of the program. Our experiment, based on Coate and Loury’s (1993) model of statistical discrimination and affirmative action, is designed to shed light on two issues: (1) whether, and in what cases, the theoretical prediction of the disincentive effects of affirmative action on human-capital investment is observed empirically to materialize; and (2) whether affirmative action needs to be a permanent intervention in order to have a lasting effect; both of these issues have obvious relevance for affirmative-action policy. Laboratory experiments are particularly well suited for examining these issues, as they allow for observation of (and indeed, control over) important variables, such as firms’ beliefs about individual workers’ productivity, that are typically unobservable in the field. Also, in lab experiments—as opposed to field experiments or field data analysis—the experimenter can easily make exogenous changes to policies (e.g., imposing or removing affirmative action), allowing for ceteris paribus comparisons. While it’s true that laboratory experiments can raise concerns of external validity (what do the decisions of undergraduate students in a computer lab tell us about the choices workers in the real world would make?), it is less of a problem in the current study, as we are interested primarily in how investment in skills changes under affirmative action, or after its removal; we are relatively less concerned about absolute levels of investment. So, even if subjects in our experiment are more—or less—likely to invest than real workers, our main results still have external validity as long as there is no reason to expect this over- or underinvestment to vary by policy treatment.

Our experimental design focuses on the behavior of individuals belonging to the historically-disadvantaged group of workers; the behavior of firms and of the advantaged group of workers are automated, and conform to the implications of Coate and Loury’s model. Subjects in the experiment, cast in the role of workers in the disadvantaged group, choose whether or not to invest in skill acquisition, then are placed by their employer into either a high-wage, high-skill task or a low-wage, low-skill one, based on a test score (which is correlated with the investment decision but not directly with group membership) and which group they come from. During the experiment, an affirmative-action policy is imposed—requiring equal proportional allocation of the two groups into the high-skill task—and in a later round the policy is removed.

The results of our experiment give mixed support to the theory. The data generally do not conform to theoretical point predictions, primarily because we observe investment levels substantially higher than predicted. It is not clear why this overinvestment (also seen in studies by Fryer et al., 2005, p. 165, and Kidd et al., 2008) occurs. One possible explanation is the Hawthorne effect (Landsberger, 1958), a psychological phenomenon in which experimental subjects improve their behavior in some way—not because of the specific manner in which the environment is changing, but simply because it is changing (or alternatively, because the subjects know that their response to the change is being observed). One piece of evidence in favor of this explanation is that changes to the policy treatment (from pre-AA to AA, or from AA to post-AA) lead to initial increases in investment in five of the six cases in which they occur, but the increase tends over time to become smaller or even disappear.11 Other possible explanations for this overinvestment include

---

11This tendency of choices in our experiment to move in the direction of the theoretical predictions is in contrast to the results of Fryer et al. (2005) and Kidd et al. (2008), both of whom find either only weak trends toward predicted behavior or no apparent trend, depending on the treatment.
optimism bias (see Armor and Taylor (2002) for some recent examples), as well as experimenter-induced demand effects.

Additionally, while we observe strong evidence that investment in skills becomes less likely as its cost increases, we do not find evidence that the population of subjects follows threshold strategies (invest if and only if the cost is below a cut-off value), contrary to the theory. On the other hand, theoretical predictions of qualitative treatment effects (either moving from one policy treatment to another or across model parameter combinations) arising from Coate and Loury’s model are largely borne out by the data. This means that despite the systematic overinvestment by workers, the sign of change in investment due to either imposing or removing an affirmative-action program is generally the same as the theoretically predicted sign.

These results lead us to some natural conjectures regarding affirmative-action policies in the outside world. The success of the theory’s directional predictions means that behavior in our experiment tended to reflect the implications of the underlying theoretical model. As a result, we conclude that the qualitative effects of introducing or repealing affirmative action are likely to be sensitive to characteristics of the labor market. In particular, in view of the fact that our experimental treatments were based on variations of the parameters $\pi_0^0$ and $\pi_0^b$, which quantify firms’ beliefs about the skills of the two worker types, we can interpret our results as implying that responses to affirmative-action policies, such as whether their effects outlast the existence of the programs themselves, will depend on the nature of the original negative stereotype held by firms—which unfortunately is unobservable outside the laboratory.

A more optimistic conjecture arises from the systematic overinvestment that was seen in the experiment. If this overinvestment carries over to the outside world, this would imply not only that changes to affirmative-action policies should have a more positive impact on targeted-group investment than that implied by the theory, but also that firms’ negative stereotypes could be eroded, even in cases where the theory implies that they should persist. This could happen if the targeted group overinvests to such an extent that they reach or surpass the frequency ($\hat{\pi}$, in our notation) at which firms optimally choose the liberal hiring standard used for the historically advantaged group, and if they persist in doing so until firms’ beliefs adjust. Assuming these conditions are met, the likelihood would be even higher that a temporary implementation of affirmative action will have a permanent beneficial effect.\footnote{Indeed, it might be that our experimental results understate the true level of overinvestment that could be expected in the outside world. While the investment decision in our experiment consisted of a zero–one choice, the corresponding decision in the outside world is perhaps better approximated by a series of real-effort tasks. The experimental literature on real-effort tasks suggests that subjects often have an intrinsic motivation to do well, resulting in more effort put into the task than would be predicted by a pure disutility-of-effort model (see, e.g., Brüggen and Strobel, 2007). Analogously, if people outside the lab have an intrinsic motivation to acquire knowledge and skills, then “overinvestment” (relative to the prediction of a pure disutility-of-effort model) might be even more prevalent than it was in our experiment.}

We close by mentioning some possible extensions to our work. One obvious set of possible extensions would examine the effects of changes to how the experiment is framed to subjects. As mentioned in Section 3.1, our description of the experimental environment deliberately avoided the use of phrasing such as “discrimination”, “stereotyping” and “affirmative action”. Most experimental economists believe that such framing can have an effect on behavior, and there is some evidence that it does.\footnote{See, for example, Sally (1995), who finds in a meta-analysis of prisoners’ dilemma experiments that messages tend to be more credible when they are described by the experimenter as “promises”, even though such wording has no formal effect on the
ings that can be aroused—across the political spectrum—by affirmative action, one might expect changes in framing to affect behavior in our experimental setting. However, due to the multiple forces affecting investment behavior, it is difficult to predict the direction these changes would have.

A second set of extensions—possibly complementary to the first set—would incorporate phenomena from the psychology–of–discrimination literature. For example, Steele and Aronson (1995) consider stereotype threat, according to which the performance of individuals, when aware that they face a negative group stereotype, is disrupted in a way that confirms the stereotype. In our experiment, stereotype threat is probably not an issue, due to our use of fairly neutral language in the instructions. However, one could imagine extensions of the Coate–Loury model in which the historically disadvantaged workers suffered from stereotype threat, in their performance on the test (so that their scores were lower, for a given level of investment, than those of the historically advantaged group), in their performance on the job (so that investment was necessary, but not sufficient, for historically disadvantaged workers to ensure qualification for Task 1), or both. To the extent that stereotype threat actually exists in populations facing negative stereotypes, such extensions would not only provide more realistic predictions, but would yield a useful model for experimental testing.

References


messages’ ability to bind the sender.

14 Also see Günther et al. (2010) for a replication of stereotype threat in an economics experiment.


