Implementation and removal of an affirmative–action quota: The impact on task assignment and 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 VIC 3800, Australia
Lata.Gangadharan@monash.edu

Michael P. Kidd*
School of Economics and Finance
Queensland University of Technology
Brisbane QLD 4000, Australia michael.kidd@qut.edu.au

October 7, 2012

Abstract

Both Canada and the United States have proactive federal legislation that attempts to address employment inequities across specific target groups, namely females, visible minorities, indigenous peoples, and the disabled. The US has a long tradition of affirmative action, dating back to President Kennedy's 1961 Executive Order No. 10965; Canada enacted the Employment Equity Act in 1986. Both countries' programs attempt to ensure the workforce is representative of the labor market as a whole. However whereas the US adopts explicit quotas, Canada relies on employer submitted equity plans involving timetables, targets and goals.

Employment Equity and Affirmative Action policy has attracted significant controversy, with several high profile court cases and the repeal of state/provincial legislation. A seminal paper by Coate and Loury (1993) examines the theoretical impact of introducing an affirmative action quota on firms' task assignment decisions and workers' skill acquisition. Unfortunately the theoretical impact of affirmative action is ambiguous. The current paper employs a laboratory experiment to shed empirical light on the theoretical ambiguity.

Journal of Economic Literature classifications: 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 and Kidd was at Deakin University. 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 terms affirmative action and employment equity are used to describe policies aimed at improving the welfare of particular target groups in the population, generally on the basis of disability, 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 Canada and the United States. [1] However, affirmative–action policies have come under attack, particularly in the US, 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 US 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.

There have also been challenges in Canada, particularly at the provincial level, with the Ontario legislation abandoned a little over a year after its enactment – a change of government being the catalyst. The Canadian legislation has a narrower focus restricted to employment equity and unlike the US excludes equity of access to education. This combined with the fact that Canada does not use explicit quotas possibly explains the somewhat lower level of disquiet compared to the US.

Given the likelihood that pressure against 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'

[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. Also, Malaysia has an affirmative action program providing preferential access to education to the native Malays.

attributes adjust over time, and thus depend on both current and past policies. Similarly the effects of repealing an affirmative–action policy cannot be assumed to be the diametric opposite of the effects of enacting such a policy.

Statistical discrimination (Phelps, 1972; Arrow, 1973) provides one promising explanation for the presence and persistence of discrimination. The underlying idea is that employers may rationally use workers' observable characteristics—such as ethnic group or sex—as a proxy for underlying 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) may become self–fulfilling. 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. 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, ceteris paribus. In this paper, we use a laboratory experiment to examine the effects of affirmative action, modeled as a simple quota. This enables us to explicitly examine the theoretical predictions of Coate and Loury (1993) within an experimental context.[3]

Specifically, we focus on individuals in the target group, who initially face a negative stereotype, and 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 arising from the stylized model in Coate and Loury's (1993) seminal paper. First, consider a prevailing negative stereotype. Coate and Loury demonstrate that, depending upon the values of parameters that characterize this stereotype, the introduction of affirmative action may 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

[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

[3] We realize this context is closer to the US legislation than the Canadian. Unfortunately given the intricacies of the Canadian intervention it would be very difficult to model explicitly. One might consider the simple quota scheme as an extreme version of the Canadian Employment Equity Act.

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. If the target group knows that they must be employed according to some legislated quota, this might lead to a disincentive to invest in skills and become qualified for the higher–level task. 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, affirmative action may require either a temporary or permanent intervention.

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 stereo- type. 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.

The contribution of our paper to the experimental literature on job market signaling and statistical dis-crimination is twofold. First and most importantly, we attempt to determine the effects of not only introducing affirmative action (as Kidd et al. (2008) 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 over investment 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.

2 Theoretical framework

Our testable hypotheses come from Coate and Loury's (1993) theoretical model of statistical discrimination and affirmative action. In this section, we summarize the essential components of the model; the interested reader can refer to Coate and Loury 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; λ 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 to either 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. Firms do not directly observe whether a particular worker is qualified for T_1. However, firms observe the worker's type, along with a test result that is correlated with the worker's investment choice, prior to the allocation of workers to tasks. This is the only choice the firm makes: it does not choose which people to hire (or fire), nor does it set wages.

Each worker decides whether to undertake a costly investment in skills, which ensures qualification for, but not assignment to, T_1. All workers then take the test 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 xq (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. 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.

Whether a worker invests depends on the cost of investment (observed prior to making the decision), and on the employer standards determining the probability of an investor versus non–investor being assigned to T_1. The employer starts with a prior belief π_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 θ, the employer forms a posterior probability ξ that depends on π_i and the likelihood ratio φ (θ) (the odds that a worker with a given score θ is unqualified). If ξ 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. 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. 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 (π_w, π_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 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 α 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 π_w^0 and π_b^0. If $\pi_w^0 > \pi_b^0$, $\pi_w^0 > 0.5$, and λ (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 $\pi_b^0 < \pi_w^0 < 0.5$, 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), our experiment is one–sided, with subjects making decisions in the role of workers, and with computerized firms. 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 should be less noisy, as we can abstract from 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 $\left(\pi_w^0, \pi_b^0\right)$. 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 $\pi_w^0 > \pi_b^0$). 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 π_w^0 and π_b^0. 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 $\left(\pi_w^0, \pi_b^0\right)$ 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 (πw , πb) are assumed to be equal to the initial parameters (π_w^0, π_b^0) from the pre–AA treatment. In later AA rounds, firms update their beliefs as follows: π_b^{t+1} is equal to the proportion of Bs investing in round t, subject to a cap of π_w^0, while $\pi_w^t = \pi_w^0$ 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 $\alpha = \frac{\pi_w^t - \pi_b^t}{1 - \pi_b^t}$, which of course will vary with π_b^t.

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 πb after the imposition of affirmative action (i.e., $1 - \pi_w^0$) is less than critical cut-off $\hat{\pi}$, 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. We vary the model parameters between–subjects, with the pairs (π_w^0, π_b^0) = (.4,.1), (.8,.1), and (.8.4); we refer to these as our model treatments. We vary the policy within - subjects (our *policy treatments*): pre–AA (before affirmative action), AA (affirmative action in place), and post–AA (after affirmative action is repealed). We will use the term "cell" to refer to a combination of model and policy treatment (e.g., the cell with (.4, .1) and post–AA). There were a total of nine experimental sessions, three of each model treatment, with ten subjects in each session.

Our first model treatment ("benign treatment") gives rise to a benign equilibrium, with πb predicted to rise from 0.1 to 0.4 upon the introduction of affirmative action, and remain at 0.4 when it is repealed. The second model treatment ("patronizing treatment") gives rise to a patronizing equilibrium; πb should rise from 0.1 in pre–AA to 0.2 in AA, but it should return to its original level of 0.1 in post–AA. The third model treatment ("worsening treatment") also leads to a patronizing equilibrium, with π_b predicted to fall from 0.4 in pre–AA to 0.2 in AA, but then rise back to 0.4 in post–AA. Table 1 summarizes the parameter combinations and theoretical predictions for each of our treatments.[4]

[4] These frequencies do not represent mixed strategies played by the B workers. Rather, they correspond to threshold investment costs, with a given worker choosing to invest if and only if his realized cost is less than the threshold.

Table 1: Treatment information

Model treatment	π_w^0	π_b^0	Predicted B invest. freq.			Predicted change B invest. freq.		Sessions/ subjects
			Pre-AA	AA	Post-AA	Pre-AA → AA	AA → post-AA	
Benign	0.4	0.1	0.1	0.4	0.4	Increase	No change	3/30
Patronizing	0.8	0.1	0.1	0.2	0.1	Increase	Decrease	3/30
Worsening	0.8	0.4	0.4	0.2	0.4	Decrease	Increase	3/30

Subjects 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 visually isolated PCs, and used pen and paper to answer some demographic questions and some risky–choice questions (the latter to provide information about 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. 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 probability of being assigned to Task 1 contingent on test performance and investment decision.

There were 37 rounds of play: 1 practice and 10 real pre–AA rounds, 1 practice and 15 real AA rounds, and 10 real post–AA rounds. Subjects were not told in advance 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. Each round was divided further into ten "periods". At the beginning of a round, subjects were informed of their investment cost, then decided whether to invest. This decision was binding for all ten periods within the round. Test results were drawn by the computer, and were i.i.d. over the ten periods within a round; the resulting allocation of the worker by the firm to Task 0 or Task 1 was also chosen independently across the ten periods. The purpose of this division of rounds into ten periods was our expectation that the larger amount of feedback from each investment choice would improve understanding of the environment and reduce decision errors.

A subject's payoff, gross of investment cost, was 150 points if assigned to Task 1 and 50 points if as- signed to Task 0; investment costs were drawn, i.i.d. for each subject and round, from a uniform distribution over $\{0, 1, 2,..., 100\}$. Once the investment choice was made, 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 payment was determined based on five randomly–chosen non–practice rounds. The payments for these rounds were summed and divided by 1000 to convert to Australian dollars. This sum was added to a participation fee of $20 and rounded to the nearest 10 cents. Overall earnings ranged from $27.50 to $42.50, for a session that lasted between 75 and 90 minutes, with at least 60 minutes devoted to the main portion of the experiment. 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.

3.2 Hypotheses

Our hypotheses 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, with significance assessed based on nonparametric tests requiring minimal distributional assumptions and based on only one observation per session.[5]5 In the second subsection, we disaggregate results partly, either by the round number (thus showing dynamics) or according to other variables. In the third subsection, we report estimates from multivariate parametric regression models, in order to disentangle the effects of our main treatment variables from those of various other variables.

4.1 Aggregate results

Aggregate investment frequencies over all non–practice rounds of each cell (combination of model treatment and policy treatment) are shown in Table 2. Also shown are the frequencies from the last three

[5] See 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.

rounds of each policy treatment—which serve as a measure of behavior after subjects have gained some experience—and the associated theoretical predictions.

Table 2: Aggregate frequencies of B group investment, all non–practice rounds

Treatment	Frequency of investment (in %)								
	pre–AA (rounds 2–11)			AA (rounds 13–27)			post–AA (rounds 28–37)		
	All rounds	Last 3 rounds	Equil. Pred.	All rounds	Last 3 rounds	Equil. Pred.	All rounds	Last 3 rounds	Equil. Pred.
Benign	45.0	43.3	10	50.9	43.3	40	48.0	46.7	40
Patronizing	41.0	31.1	10	35.6	26.7	20	31.0	27.8	10
Worsening	59.0	63.3	40	45.8	37.8	20	48.3	50.0	40

Two features of these data stand out. First, there is substantial overinvestment; in each cell, the observed frequency of investment is higher than the theoretical prediction, and the difference varies from 8–35 percentage points. If we pool the 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 tend to qualitatively reflect 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, $U` = -2.91$, $p < 0.05$), matching the qualitative difference in their predicted frequencies of 0.4 versus 0.1. In the AA stage, investment is higher in the benign treatment than in the other two, qualitatively consistent with the predictions of 0.4 versus 0.2, though this observed difference is statistically insignificant (pooled patronizing and worsening treatments, $U` = -1.50$, $p > 0.10$). In the post–AA stage, investment in the patronizing treatment is significantly less than in the other two (pooled benign and worsening treatments, $U` = +\infty$, $p \approx 0.012$), again mirroring 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 (as dark circles), along with the theoretical predictions (open circles) and investment frequencies for the simulated W workers, which by construction are equal to their theoretically predicted values. Nearly always, observed frequencies begin well above their predicted levels. They trend downward over time, but typically remain above the predictions, even in the last round of a policy treatment.

Figure 1: Frequency of B–type workers' investment, benign treatment, all rounds

(*Note: Rounds 1 and 12 were practice rounds*)

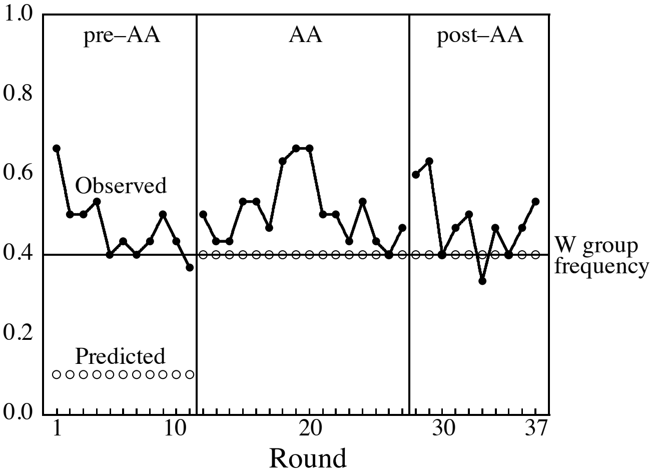
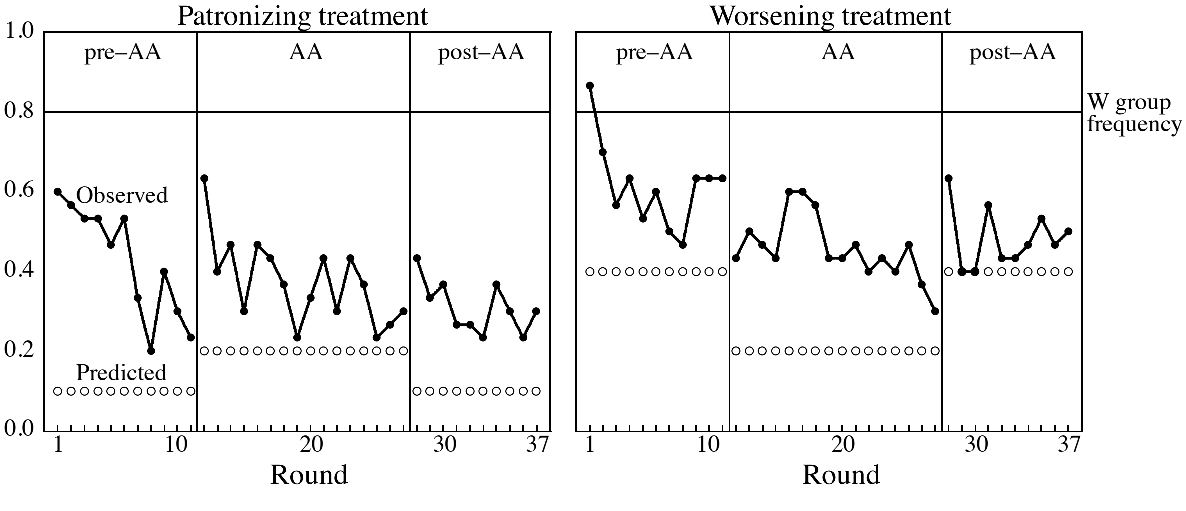


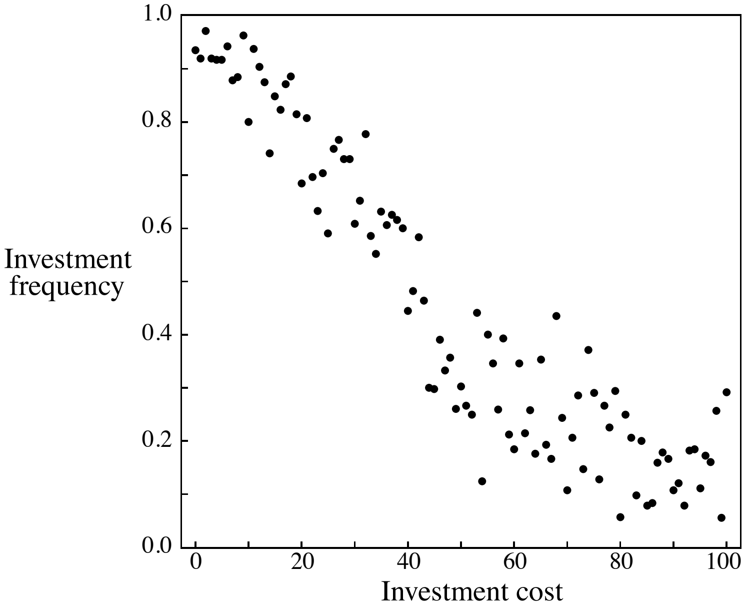
Figure 2: Frequency of B–type workers' investment, patronizing and worsening treatments, all rounds (*Note: Rounds 1 and 12 were practice rounds*)



We can also see the effects of introducing and subsequently removing affirmative action. In all three model treatments, introducing affirmative action leads to an apparent change in B workers' 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 (the theory implies no effect in the benign treatment and a decrease in the patronizing treatment).

Figure 3: Scatter–plot of investment cost and investment frequency,

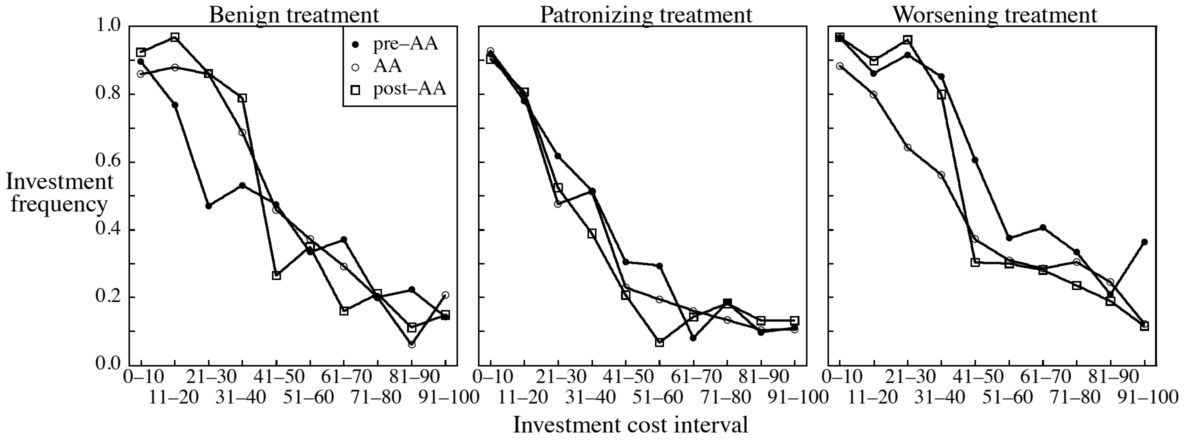
all cells, all non–practice rounds



These aggregate data disguise a fine sensitivity of B workers' 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 proportion of times in the experiment 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 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: [0, 10], [11, 20], [21, 30], and so on. The negative relationship 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 seen 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.

Figure 4: Investment frequency (by cost), disaggregated by cell, all non–practice rounds



4.3 Parametric statistics

We next report the results of probit regressions. For each model treatment (benign, patronizing, worsening), we estimate a restricted model looking for average effects on investment, and an unrestricted model intended to detect time–varying effects. In each, the dependent variable is an indicator for an "invest" choice. Right– hand–side variables are a constant, the realized cost of investment, and indicators 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.[6] All of the models were estimated using Stata (version 10), and incorporated individual–subject random effects. Table 3 shows the usual coefficient estimates and standard errors, along with p–values corresponding to tests of joint significance for the two pre–AA variables and for the two post–AA variables.

[6] We 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. Our sample consists of 49% male subjects; average age of the subjects is 23 years; approximately 45% were studying for a degree in commerce; 25% of the sample were in their 4th year of study or in a post-graduate program and about 52% were risk–averse according to their lottery choices. The main results were not qualitatively affected by inclusion of these variables, so for space reasons we do not report them here.

Table 3: Regression coefficients, standard errors, and significance results, non–practice rounds

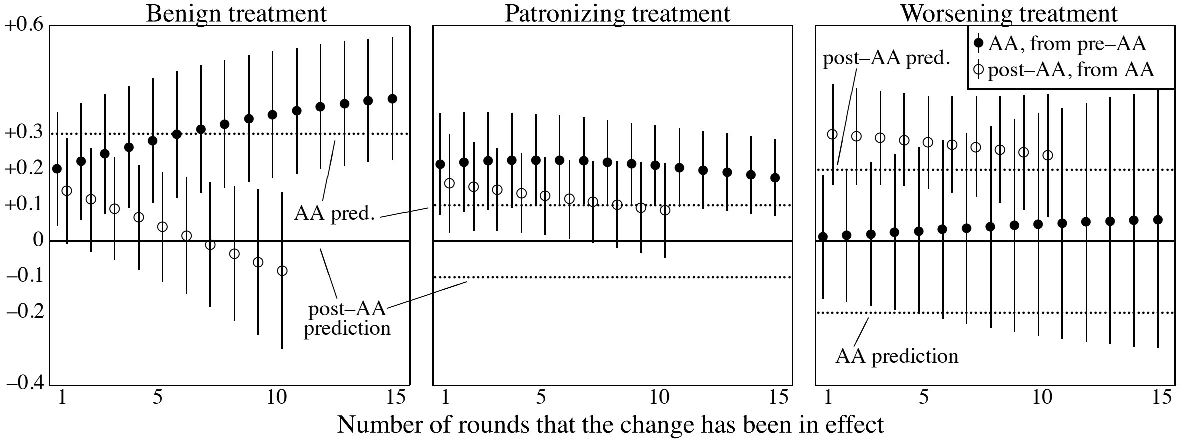
Variable	Benign		Patronizing		Worsening constant	
constant	1.633***	1.752***	1.328***	2.088***	1.423***	2.318***
	(0.173)	(0.375)	(0.220)	(0.435)	(0.183)	(0.388)
Investment cost	−0.034***	−0.035***	−0.039***	−0.038***	−0.031***	
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	
Pre-AA	–0.039	0.314	0.159	–0.061	0.610***	0.104
	(0.116)	(0.409)	(0.125)	(0.447)	(0.118)	(0.405)
Post-AA	0.013	2.129*	–0.028	0.892	0.189	0.970
	(0.117)	(1.099)	(0.129)	(1.202)	(0.115)	(1.086)
round		–0.005		−0.039**		−0.043***
		(0.017)		(0.019)		(0.017)
Pre-AA * round		–0.065		–0.048		–0.010
		(0.036)		(0.039)		(0.036)
Post-AA * round		–0.063		–0.014		–0.007
		(0.036)		(0.040)		(0.036)
Joint test, pre-AA vars.		p ≈ .09		p ≈ .06		p ≈ .96
Joint test, post-AA vars.		p ≈ .12		p ≈ .12		p ≈ .003
Pseudo–R^2	0.295	0.302	0.336	0.347	0.274	0.283

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

Some evidence that investment does vary across policy treatments comes from these joint tests: 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 comes from estimates of the overall incremental effects of our pre–AA and post–AA treatments on investment (i.e., through both the value of the relevant indicator variable and its interaction with the round number). These estimated values, along with 90% confidence intervals, are shown for each round in Figure 5.[7]7 Also shown in each of the three panels is the horizontal segment corresponding to a zero effect, and dotted segments showing the theoretically predicted effects (as shown in Table 1). In each panel, the first dark circle shows the initial effect of affirmative action on B workers' investment, compared to the hypothetical situation where it had not been introduced (i.e., with pre– AA continuing for another round). The trajectory of dark circles from round to round shows how this effect changes with time. Similarly, the first open circle shows the initial effect of repealing affirmative action— compared to the hypothetical case where affirmative action continued to be in place—and the trajectory of open circles shows how this effect changes from round to round.

[7] Note 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.

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)



Thus, in the benign treatment, the dark circles in Figure 5 indicate a significant and persistent positive effect on Bs' investment from introducing affirmative action, while the open circles show a small, insignificant initial increase in investment due to removing affirmative action, which declines over time (though remaining insignificant). Both 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 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 our hypotheses, though it is a weak implication of the theory. (The model makes a stronger prediction not observed in the data: that there is a cell–specific threshold value for the investment cost, such that the worker invests for sure if the actual cost is lower, and does not invest if the 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 in the disaggregated data (Figure 4), and the significant negative coefficients associated with the investment cost variable, seen in the regressions (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 (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 grows over time. Removing affirmative action, by contrast, leads to a positive but insignificant change initially, and while this incremental effect decreases over time, 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 round-by-round descriptive statistics support the result visually (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 predicts 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 (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

Changes in the political and legal climate in some countries, and the theoretical possibility that affirmative action policies may lead to a worsening of the original negative stereotype, give reason to believe that the future of these policies is in some doubt. As regional and national governments consider dismantling affirmative–action programs, it becomes increasingly important to understand the likely effects of such changes.[8]

This paper presents the results from a novel experiment that enables a direct comparison of labor–market outcomes before and after implementation of an affirmative–action program, and also after a subsequent repeal of the program. Our experiment is based on Coate and Loury's (1993) model of statistical discrimination and affirmative action. We examine (1) whether the theoretical prediction of the disincentive effects of affirmative action is observed to materialize, and (2) whether affirmative action needs to be a permanent intervention in order to have a lasting effect, both of which have obvious policy relevance. Laboratory experiments are particularly well suited for examining these issues, as they allow observation of important variables that are typically unobservable in the field, and exogenous variation of policies that are difficult to change in the outside world. While it is true that laboratory experiments can raise concerns of external validity, this is less of an issue in the current study, as we are interested primarily in how investment changes under affirmative action, or after its removal, as opposed to absolute levels of investment. So, even if our subjects 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.

[8] 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 US 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).

Our results give mixed support to the theory. Investment levels are typically substantially higher than their corresponding theoretical point predictions, as has been seen in previous studies (e.g., Fryer et al., 2005, p. 165; Kidd et al., 2008). One possible explanation for this overinvestment 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 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 increases tend to shrink or even disappear over time.[9] Other explanations include optimism bias (Armor and Taylor, 2002), as well as experimenter–induced demand effects.

On the other hand, theoretical predictions of *qualitative* treatment effects (across either model or policy treatments) arising from Coate and Loury's model are largely borne out by the data. This means that de- spite 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 reflects 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 design was based on variation of firms' beliefs about the skills of the two worker types, our results imply that responses to affirmative–action policies will depend on the nature of firms' negative stereotypes—which unfortunately is unobservable outside the laboratory.

A more optimistic conjecture arises from the systematic overinvestment that we observed. If this over- investment carries over to the outside world, this would imply not only that policy changes should have a more positive impact on targeted–group investment than that implied by the theory, but also those 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 the frequency at which firms optimally choose the liberal hiring standard used for the historically

[9] This tendency of choices in our experiment to move in the direction of the theoretical predictions is in contrast to Fryer et al. (2005) and Kidd et al. (2008), both of whom find either only weak trends toward predicted behavior or no apparent trend.

advantaged group, until firms' beliefs adjust. If so, it would be even more likely that a temporary implementation of affirmative action will have a permanent beneficial effect.[10]

We close by mentioning some possible extensions to our work. One obvious set would examine changes in how the experiment is framed. Our instructions to subjects avoided the use of phrasing such as "stereo- typing", "affirmative action", etc. Most experimental economists believe that framing can have an effect on behavior, and there is some evidence that it does.[11]11 Given the intense feelings 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 complementary, set of extensions 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.[12]12 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 their investment was necessary, but not sufficient, 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.

While this paper focuses on discrimination of workers by firms, in many organisations one also observes discrimination by peers. This would imply that B–type workers prefer working with other B's and W's with other W's. This could have in–group and out–group implications that would be interesting to investigate. In addition, in the context of gender discrimination, it is often argued that female success

[10] Indeed, it might be that our 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 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" might be even more prevalent than it was in our experiment.

[11] 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 messages' ability to bind the sender.

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

is seen to be a signal of hard work, whereas male success is taken to be a signal of brilliance. Hence, the relationship between investments made by workers and the signals given from outcomes could differ according to worker type, leading to discrimination in promotion to higher–level jobs. These are exciting areas for future research.

References

Armor, D.A. and S.E. Taylor (2002), "When predictions fail: the dilemma of unrealistic optimism", in *Heuristics and Biases: The Psychology of Intuitive Judgment*, T. Gilovich, D. Griffin, and D. Kahneman, eds., Cambridge University Press, Cambridge, pp. 334–347.

Arrow, K.J. (1973), "The theory of discrimination", in *Discrimination in Labor Markets*, O. Ashenfelter and A. Rees, eds., Princeton University Press, Princeton, NJ, pp. 3–33.

Brüggen, A. and M. Strobel (2007), "Real effort versus chosen effort in experiments", *Economics Letters* 96, pp. 232–236.

Coate, S. and G.C. Loury (1993), "Will affirmative–action policies eliminate negative stereotypes?" *American Economic Review* 83, pp. 1220–1240.

Feltovich, N. (2005), "Critical values for the robust rank–order test", *Communications in Statistics— Simulation and Computation* 34, pp. 525–547.

Feltovich, N. and C. Papageorgiou (2004), "An experimental study of statistical discrimination by employers", *Southern Economic Journal* 70, pp. 837–849.

Fryer Jr., R.G., J.K. Goeree, and C.A. Holt (2005), "Classroom games: experience–based discrimination", *Journal of Economic Education* 36, pp. 160–170.

Fryer Jr., R.G., G.C. Loury, and T. Yuret (2008), "An economic analysis of color–blind affirmative action", *Journal of Law, Economics, and Organization* 24, pp. 319–355.

Günther, C., N.A. Ekinci, C. Schwieren, and M. Strobel (2010), "Women can't jump? An experiment on competitive attitudes and stereotype threat", *Journal of Economic Behavior & Organization* 75, pp. 395–401.

Kidd, M.P., P. Carlin, and J. Pot (2008), "Experimenting with affirmative action: the Coate and Loury model", *The Economic Record* 84, pp. 322–337.

Landsberger, H.A. (1958), *Hawthorne Revisited*, Cornell University Press, Ithaca, NY.

Myers, C.K. (2007), "A cure for discrimination? Affirmative action and the case of California's Proposition 209", *Industrial and Labor Relations Review* 60, pp. 379–396.

Phelps, E.S. (1972), "The statistical theory of racism and sexism", *American Economic Review* 62, pp. 659–661.

Ray, D. and R. Sethi (2010), "A remark on color–blind affirmative action", *Journal of Public Economic Theory* 12, pp. 399–406.

Sally, D. (1995), "Conversation and cooperation in social dilemmas: a meta–analysis of experiments from 1958 to 1992", *Rationality and Society* 7, pp. 58–92.

Siegel, S. and N.J. Castellan, Jr. (1988), *Nonparametric Statistics for the Behavioral Sciences*, McGraw– Hill, New York.

Steele, C.M. and J. Aronson (1995), "Stereotype threat and the intellectual test performance of African Americans", *Journal of Personality and Social Psychology* 96, pp. 797–811.