Some Experimental Evidence on the Coate and Loury Model of Affirmative Action

P Carlin (IUPUI) M Kidd (Deakin) and Jay Pot

Abstract

Coate and Loury (1993) suggest the impact of affirmative action on a negative stereotype is theoretically ambiguous leading to either: a benign equilibrium in which affirmative action eradicates the negative stereotype leading to equal proportional representation of the two groups; or alternatively a patronising equilibrium in which the stereotype persists. The current paper examines this theoretical ambiguity within the context of an experimental design. Although benign and patronising equilibria are equally plausible in theory, the experiments easily replicate the benign equilibrium, but find that a certainty effect leads to over-investment and divergence from the theoretically predicted patronising equilibrium.

JEL C92, J15, J71.

Affirmative action is a controversial and wide-ranging policy issue in Australia and elsewhere. Both interpretation and implementation vary across country. In the US affirmative action is sometimes interpreted as a quota system to correct prior injustice. In Australia the Equal Opportunity for Women in the Workplace Act 1999 replaced the prior Affirmative Action Act (1986). The Act aims to ensure recruitment, selection and promotion are based upon merit without regard to gender. This would appear to rule out corrections for previous discriminatory practices. However, it does allow for a concerted effort to expand the scope of job applicants in order to ensure no group is excluded on the basis of either past or present discrimination.

One recent well advertised Australian flirtation with AA is the Australian Labor Party commitment that by 2002 a minimum of 35% of ALP candidates for winnable seats would be occupied by women, a target raised to 40% in October 2002. One other recent example is the controversy raised by the Sex Discrimination Act (Teaching Profession) Bill 2004-5. The bill sought to amend the Act to provide special dispensation in the area of education, allowing the provision of teaching scholarships to males only in order to address the gender imbalance in teaching. The bill passed through the House of Representatives, but stalled in the Senate.

In the US the legality of race-based quota schemes regulating entry to Law school have recently come under challenge. In November 2006, Michigan became the third state in the US to ban affirmative action in state-supported university admissions. On the other hand, Wayne State University, located in Michigan, announced it will still consider family background and social/economic disadvantage to be relevant criteria in their admissions process.

India has a long tradition of affirmative action policy, reserving almost a quarter of government jobs and places in state-funded colleges for individuals of the lowest caste. A bill introduced in December 2006 further strengthened the policy, increasing the scope to include other low caste groups and the 'quota' to nearly 50 percent. Violent disturbances within immigrant neighbourhoods of Paris in summer 2006 raised the awareness of inequality. This stimulated European interest in affirmative action type policies which might increase access of young immigrants to permanent jobs.

The seminal work of Coate and Loury (1993) (henceforth CL) suggests that the impact of affirmative action (AA) upon a negative stereotype, i.e. a belief that one group is less productive than another, is theoretically ambiguous. The model demonstrates that AA will lead to either: a benign equilibrium in which affirmative action eradicates the negative stereotype, leading to equal proportional representation of the two groups; or alternatively a patronising equilibrium in which the stereotype persists.[1] Given the topical nature and policy import of AA this theoretical ambiguity requires empirical investigation; a non-trivial task as stereotypes are not directly observed. Following the literature on job market signalling and screening we adopt an experimental approach.[2]

The focus of our current analysis is two-fold: Can we replicate the 'stickiness' of stereotypes in the absence of AA and the theoretical ambiguity of the impact of AA within the context of an experimental design? Secondly, although benign and patronising equilibria are equally plausible in theory – is empirical reality equally agnostic? In brief we find that in order to replicate the 'stickiness' of negative stereotypes within an experimental context, CL's belief updating mechanism must be modified. In terms of the

impact of AA, we readily replicate the theoretically predicted benign equilibrium, but cannot attain the precise patronising equilibrium suggested by CL.

I. Brief review of existing experimental literature

There are a number of papers which examine signalling (Spence (1973)) within an experimental context. See for example Kübler, Müller and Normann (2005); Fryer, Goeree, and Holt (2005) (henceforth FGH), Anderson, Fryer and Holt (2002) and Holt (2007); and Feltovich and Papageorgiou (2004) (henceforth FP). The two sets of papers most directly relevant to the current analysis are the paper by FGH and associated papers involving Holt, together with the paper by FP.

To our knowledge FGH and accompanying papers are the only prior experimental work directly linked to the theoretical analysis of Coate and Loury (1993). The focus of the FGH paper is on whether the existence of a negative stereotype can be replicated in an experimental setting. The experiment set up is rather ingenious and tightly linked to the underlying CL structure. The experiment is two-sided, with green and purple workers randomly allocated across firms. Workers decide whether or not to invest, and firms whether or not to hire the worker. Investment on the part of workers increases the performance on a test (acting as a noisy signal) which firms in turn use to make hiring decisions.

The FGH experimental results are interesting. Their results (Figure 1, pg. 164) demonstrate a clear separation between groups, with 80-90 (40-50) percent of green (purple) workers investing. This is despite the fact that both groups initially face ex-ante identical trade-offs, i.e. the distribution of costs of investment are identical.[3] This investment pattern leads to just over 80% of green workers being hired and 35-55% of

purple. In the FGH set-up at the end of each round, summary public information is revealed, i.e. the historical proportion of each group (green/purple workers) investing and the proportion of each group hired. The public revelations appear to help maintain the observed separation of groups and associated negative stereotype for purple workers.

Interestingly in their Figure 3, FGH attempt to encourage group separation by having purple (green) workers initially draw exclusively from the top (lower) half of the cost distribution. This initial cost difference is an attempt to create an environment in which purple workers start at a disadvantage and are thus expected to invest at a lower rate, with associated hiring implications. Once the cost discrepancy disappears in future rounds, the question is whether the investment and hiring separation between green and purple workers is maintained. The results are somewhat mixed; in FGH, counter to intuition, there is little initial investment separation despite the differential investment cost profiles. Intriguingly after investment cost distributions are equalised, the previously disadvantaged group invests at a higher rate than the advantaged group. The purple group end up being hired at a much greater frequency then the green group. Anderson, Fryer and Holt (2002) and Holt (2007) report similar experiments where, instead, the initially disadvantaged group continues to be ‘disadvantaged’ despite the equality of green/purple cost distributions.

The main message of this group of papers is that one can replicate, within an experiment, the notion of a negative stereotype underlying the CL analysis**.** However, the papers shed little light on the question of <u>when</u> a historical disadvantage will carry over to the future, reinforcing the negative stereotype and when the initial disadvantage will be

overcome. To some extent the issue revolves around the stickiness of firm beliefs about the investment behaviour of worker groups.

Feltovich and Papageorgiou (2004) (FP) examine the issue of statistical discrimination and employer behaviour and beliefs in a novel one-sided experiment design. In essence the focus is on how sensitive historical firm beliefs are when confronted with new information. Experimental participants (firms) initially find evidence that one pool of workers is more productive than the other. New information is revealed such that the prior information on relative productivity is potentially no longer informative. The question is to what extent the priors carry forward? In summary their evidence suggests that strong priors are difficult to simulate in the experimental context; firms appear to learn quickly and update beliefs accordingly. Their evidence suggests that negative stereotypes are thus likely to disappear.

The major novelty of our analysis is that we attempt to examine the impact of affirmative action. None of the previous cited literature focuses on that issue. The paper shares the tight link with CL underlying FGH, but as with FP is limited to a one-sided experiment. Because the interaction between workers and firms is complex we choose to focus on the simplest structure—workers playing the role of experimental participants and the employer side simulated by a computer.[4] Apart from avoiding complexity per se there are two main reasons why we focus on the one-sided experiment. CL start from a negative stereotype, i.e. an initial set of firm beliefs *re* group productivity and distinguish between two possible equilibria for the impact of affirmative action. In the context of the CL special case, the focus of our experiment, whether affirmative action leads to the benign or the patronising equilibrium depends crucially on the initial firm beliefs. The

beauty of the FGH approach is that by having a two-sided experiment firms make decisions based on whatever their beliefs are. The downside is they cannot impose a particular set of beliefs in order to examine the predictions of the impact of AA. We use the first set of experimental rounds in order to create an environment in which a negative stereotype exists and examine worker reaction. By tracking worker investment behaviour we can also examine the sensitivity of the theoretically imposed initial firm belief to various hypothetical methods of updating beliefs.

The downside of a one-sided experiment is one loses the richness of agent interaction. In the FGH experiment workers need to discern how firms will treat them. The literature suggests that experimental agent behaviour differs depending upon whether they are interacting with other agents or with a robot.[5] [See Calegari et al (1998), Cason and Friedman (1999), Datta (2005), and Cason and Sharma (2006).] This literature investigates whether experiments where human subjects interact with robots (programmed response) rather than each other affect the ability of experiments to converge to a theoretically predicted equilibrium. The first three papers listed above confirm the proposition that the use of robots leads to more frequent verification of theoretical equilibria in settings as diverse as posted offer markets with buyers and sellers, price-setting, differentiated duopoly, and auditor/client interactions on price and independence. Datta (2005) suggests that experimental markets with human subjects on both sides have difficulty reaching the predicted equilibrium because human buyers may act strategically and search sub-optimally. Cason and Sharma (CS hereafter) report on an experiment where they try to implement a correlated equilibrium. As in our experiment, CS comment that the predicted equilibrium assumes optimal decisions on both sides of

the market. CS find that the correlated equilibrium is achieved when robots are used on one side of the market but not when both sides are human. They conclude that equilibrium is not achieved in the latter case because doubts about the rationality of other players make it sub-optimal to follow equilibrium strategies that are optimal only if both players are choosing optimally. Hence a one-sided experiment may increase the likelihood of replicating the CL equilibrium predictions.

II. Coate and Loury's (1993) model

We first briefly consider the general model of CL and then the special case upon which our experiment is closely based. Consider a labour market with a large number of identical firms and workers who are ex-ante identical apart from racial identity. A representative firm is randomly matched with a set of workers who are in turn allocated to one of two tasks, *Task 0* (T_0), or *Task 1* (T_1). Workers who have invested in the requisite skills are qualified for the higher level task (T_1). Given asymmetric information the firm cannot directly discern an individual's investment choice. The firm adopts a 'noisy' test score together with, possibly, racial identity as a proxy for expected productivity and hence for allocation of workers between tasks.

Each worker decides whether or not to undertake a costly investment in skills, which ensures qualification *for*, but not allocation *to* T_1. Workers obtain a test score, $\theta \in [0, 1]$ where higher values are more likely if the worker is qualified. In particular, if $f_u(\theta)$ and $f_q(\theta)$ are the density functions of an unqualified and qualified worker respectively, the likelihood ratio $\varphi(\theta)$ (i.e. the odds that a worker with a given score, θ, is unqualified) is assumed non-increasing on [0, 1]. The employer realises a positive x_q

(negative $-x_u$) net return from assigning a qualified (unqualified) worker to T_1. Let r be the ratio of net gain to net loss for the firm. Finally, assume that workers assigned to T_1 (T_0) receive a gross return of $\omega(0)$. The net return may differ since individuals who invest must pay cost c.

Prior to being matched with an employer workers must decide whether or not to invest. Their choice depends on the probability of an investor versus non-investor being assigned to T_1, which depends in turn on the employer standards they anticipate facing. The employer starts with a prior belief, π_i, i= b, w about the probability of a worker in the given group being qualified. After observing the worker's signal, θ, the employer forms a posterior probability, ξ, which depends on the initial value of π and the likelihood ratio, φ. CL show that it is optimal for the employer to adopt a standard

$$s^*(\pi_i) \equiv \min\left\{\theta \in [0,1] \mid r \geq \left[(1-\pi_i)/\pi_i\right]\varphi(\theta)\right\},\ i = b, w \qquad (1)$$

and assign a worker to T_1 if and only if the worker's signal exceeds $s^*(\pi_i)$. They also show that $s^*(\cdot)$ is decreasing in π, which implies that the more optimistic the employer is about a group, the easier the chosen standard.

Figure 1 characterises the CL general case. The horizontal axis is the employer standard and the vertical axis is the employer belief and/or the proportion of workers investing. The optimal locus of employer standards, given any employer belief π, is given by EE and the optimal locus of worker investment proportions, given any employer standard, s, is given by WW. The intersection represents an equilibrium whereby the employer's belief and the worker's optimal investment proportion are equal.

[Place Figure 1 about here.]

Multiple equilibria are possible including negative stereotyping equilibria such as the pair (s_w, π_w) and (s_b, π_b)[6]. Small errors of perception by the employer dissipate via a simple belief adjustment mechanism following a cobweb or stair step-type convergence to the equilibrium.

CL introduce a simple affirmative action policy that requires employers to assign the same proportion of workers from each group to T_1.[7] Under this policy, CL identify sufficient conditions for the initial negative stereotyping equilibrium to be, in their terminology, *benign*. In this case, policy leads to homogeneous employer beliefs about the two groups and thus affirmative action is required only as a temporary intervention. The alternative to the benign equilibrium is labelled *patronising* where heterogeneous beliefs about the two groups are maintained.

The special case

CL introduce a special case of their model that forms the basis for our experimental design. A number of simplifying assumptions are made. Investment costs are distributed uniformly on [0, 1]. If workers invest (are qualified), their test score, θ, is uniformly distributed over the closed interval, $[\theta_q, 1]$, and if unqualified is uniformly distributed over $[0, \theta_u]$, with $\theta_q < \theta_u$. (See Figure 2.) In the special case in the absence of

[Place Figure 2 about here.]

affirmative action, the employer will assign all (no) workers with scores above θ_u (below θ_q) to T_1. The task assignment of a worker with an uncertain test score, i.e. between θ_u and θ_q, depends upon the employer prior beliefs, π_t i = b, w. The probability that an unqualified (qualified) worker will receive an uncertain score is $p_u = f_u = \frac{\theta_u - \theta_q}{\theta_u}$

$\left(p_q = f_q = \frac{\theta_u - \theta_q}{1 - \theta_q} \right)$ so that the likelihood ratio, $\varphi = p_u/p_q$. CL show that the employer will assign a worker with an uncertain score to T_1 if $\pi \geq \frac{\varphi}{r + \varphi} \equiv \hat{\pi}$. As φ depends on the exogenously set θ's and r depends on the exogenously determined x_q and x_u, $\hat{\pi}$ is exogenously fixed. So the employer either assigns all members of a group with uncertain test scores (a liberal standard) or none of them (a conservative standard)—these are the optimal standards set by the firm in the special case.

Assuming risk neutrality, workers invest if their expected payoff net of costs, $\omega(1 - p_u) - c$ for *W*'s and $\omega(1 - p_q) - c$ for *B*'s, is greater than zero. When workers see a liberal (conservative) 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 proportion of *W* and *B* workers investing.

There is an important distinction between the general formulation of the CL model and the special case. In the CL general model there is a unique correspondence between π (i.e. the firm belief) and the cut-off value of the test score θ representing the minimum standard for assignment to Task 1. Thus a change in π_i implies a change in the optimal cut-off θ for group i and consequently the firm standard. In the special case the cut-off values of the test score i.e. θ_q and θ_u are (as explained above) exogenously set, rather than endogenously determined. Thus in the context of Figure 1 in the special case, there is no equivalent to the EE locus for the firm. Given an exogenously set θ_q and θ_u, there is exactly one pair of beliefs, π_i, i = w, b that constitutes an equilibrium. This can

be most easily seen by reference to the relationship between the exogenously given test score cut-offs and the employer beliefs in a given equilibrium:

$$\theta_q = \pi_w \theta_u \tag{2}$$

$$\theta_u = \frac{1-\pi_b}{1-\pi_w \pi_b} \tag{3}$$

In the equilibrium both equations must hold simultaneously and thus for given θ_q and θ_u there is a unique pair π_w and π_b. Convergence to equilibrium is achieved by workers adjusting their (non-optimal) investment behaviour towards the optimal level and consequently firm beliefs back to the initial equilibrium.[8] In Figure 3 in equilibrium,

[Place Figure 3 about here.]

proportion π^{0}_{b} of B workers will invest, with the same proportion gaining a Pass on the qualifying test, and thus $(\pi^{0}_{b})^2$ being allocated to Task 1 pre-AA.

Figure 3 also demonstrates the outcome for B workers after the introduction of affirmative action (mandating that equal proportions of *B*'s and *W*'s be assigned to T_1). In the AA equilibrium the employer optimally assigns all *B*'s with passing and uncertain scores and a fraction, α, of those with failing scores. In the special case, CL demonstrate that there are two types of equilibria. If $\pi_w > \pi_b$, $\pi_w > 0.5$, and λ, the proportion of *W* workers, is sufficiently large there is a unique stable equilibrium (labelled *patronising*) in which employers continue to possess negative stereotypes about *B* workers—with equilibrium $\pi_b = 1 - \pi_w < \pi_w$. Alternatively if $\pi_w < 0.5$, there may be a locally stable equilibrium where negative employer stereotypes are eliminated, $\pi_b = \pi_w$ (labelled *benign*).

III. From Theory to Experiment

The focus of our experimental analysis is clear- we wish to examine whether we can replicate the two varieties of AA equilibria characterised by CL as benign and patronising. In the CL special case the distinction is extremely clear, starting from a negative stereotype equilibrium with $\pi_w > 0.5$ ($\pi_w < 0.5$), introduction of AA should lead theoretically to the π_b = 1- π_w patronising ($\pi_b = \pi_w$ benign) equilibrium. The Flow Chart in Figure 4 characterises the basic elements of the one-sided experiment, with experimental participants playing the role of workers, and the computer the role of the firm. In the initial formulation, W workers are assumed to behave as in the CL special case, i.e. investing at the rate that precisely matches and thus confirms firm beliefs. Our focus is upon the B group and their investment behaviour.[9]

We also provide evidence on a secondary issue, the updating of firm beliefs *re* the proportion of B workers who are qualified. In Figure 4 this is represented by the middle bottom panel. The CL theoretical predictions relate to the impact of AA on a given negative stereotype. In the experimental context we attempt to replicate the existence of a negative stereotype equilibrium by holding firm 'beliefs' fixed, thus in pre-AA experimental rounds regardless of the B group's aggregate behaviour, hiring standards are unchanged. Despite this assumption we track firm 'beliefs' under two hypothetical updating mechanisms. In the first we allow firms to update ala CL, i.e. firm belief $(\pi_b)_{t+1}$= proportion of B investors in period t. Given that many of the investing B workers in pre-AA will not be allocated to Task1 we also examine an alternative updating mechanism which errs on the side of minimum adjustment, i.e. so long as what is observed is not incompatible with the belief, the belief is unaltered. The alternative assumes only the

proportion of investors assigned to T_1 together with aggregate test score information is revealed.[10] Both updating schemes are based on Bayesian updating (see Cox et al [2001]) but differ in their assumptions about the information available to the firm. Tracking beliefs in this manner sheds light on whether a negative stereotype is likely to be maintained given alternative updating mechanisms.

IV. The Experimental Design.

Experimental subjects were recruited from the undergraduate and graduate student bodies of Curtin University, Perth and Deakin University, Melbourne, Australia. In total 110 subjects participated. Although the experiments were widely advertised on bulletin boards and electronically throughout the universities, the majority of participants were in business, management or economics. The results reported in the paper arise from 11 sessions with 10 participants each. Approximately 13 students were signed up for each experimental session; in all cases at least 10 students showed up. When more than 10 showed up, the excess was randomly discarded. Students who did not participate in the experimental rounds were paid a participatory fee of $10 (Aus). Students who participated received an initial fee of $15, plus they could earn/lose up to an additional $6 depending upon their experimental decisions and partly upon luck, thus overall they earned between $9 and $21. The payment scheme was explained to the participants at the beginning of the session. No subject participated in more than one session.

Experimental materials (including items such as general instructions, information cards, the payment scheme and the set of questions about background and risk profile) are contained in the appendix [A more detailed supplementary appendix is also available

upon request]. Much of this information is also available directly from the accompanying website, see http://lab.ohwhoknows.com/admin.

Participants were seated at individual computer terminals and could not observe others' screens. Their computer screens displayed the general set of instructions which were also read aloud by the experimenter. The participants were informed that in general terms the experiment related to making decisions under uncertainty. In brief they face an employer (computer) who must decide to which of two tasks they should be assigned. Task 0 is a low level task, Task 1 can only be successfully carried out by a worker who chooses to invest in skills (i.e. is qualified). There is asymmetric information in that the firm cannot observe whether or not the worker is qualified. Each worker is given a hypothetical test, the results of which help determine the firm's assignment decisions. The investment is costly (with the actual cost revealed prior to the investment decision) but increases the likelihood of being assigned to task 1 and thus receiving a higher wage. Participants are informed that the amount of money they earn depends upon the decisions they make and partly upon luck. Further details are provided below.

In each round of the experiment the computer screen displays the rule the firm uses (hiring standard) to assign workers between the two tasks, e.g. 'assign all applicants with clearly passing scores to Task 1 and those with uncertain or failing scores to Task 0'. The probability of achieving each summary test grade (pass/uncertain/fail) and associated probability of T_1 assignment is documented as well as the cost of investing. To start with, the firm rule is based upon the chosen initial beliefs and apart from random changes to the cost of investing the virtual card is unaltered during the pre-AA rounds. After an initial practice round followed by a set of 10 pre-AA rounds, participants are

warned that the firm hiring standard may change, and that there will be a further practice round. The total number of rounds to be played is not revealed. In fact there are 15 AA rounds.

Once the investment choice is made in a given round, the participant receives information about their assignment and their payoff in points. They are told this is private information that should not be shared with anyone else. Each participant is told they will go through several rounds and, at the end of the experiment, three rounds will be randomly selected, points summed and divided by 50 to convert to dollars. Total payment was thus \$15 plus/minus the above sum. Participants were not provided with any information about the negative stereotyping environment or the introduction of affirmative action.[11] The intent is that participants are simply responding to the incentives they see. Subjects performed the experiment only once with one set of parameters so there are no carryover effects from previous experimental sessions or earlier sets of parameters.

Table 1 summarises the basic parameter structure of the experiments. Recall the theoretical predictions underlying the CL special case: with $\pi_w > 0.5$ ($\pi_w < 0.5$) introducing AA leads to the patronising (benign) equlibria $\pi_b = 1 - \pi_w$ and ($\pi_b = \pi_w$) respectively. The first three sessions listed in Table 1 set $\pi_w < 0.5$ and examine whether the B group invests such that π_b converges to the theoretically predicted benign equilibrium. The next three experiments select $\pi_w > 0.5$ and examine whether the B group invests such that π_b converges to the theoretically predicted patronising equilibrium, i.e. $\pi_b = 1 - \pi_w$. Experimental sessions 7 through 11 explore a number of robustness checks. Session 7 provides a robustness check, selecting an alternative

combination (π_w, π_b); with $\pi_w > 0.5$ once again the prediction is convergence to the patronising equilibrium after implementation of AA. Session number 8 considers the effect of allowing π_w to vary from round to round based upon a prior set of experimental results for the W group; sessions 9 through 11 check the robustness of earlier results to an alternative framing of the choice scenario to avoid any possible tendency of student subjects to over-invest, simply because of the framing of the choice, when faced with a test scenario with pass/fail results.

IV. Experimental Results

Table 2 presents the summary information on risk attitude, based on the Holt and Laury (2002) analysis. (The detailed questions are in a supplementary appendix available from authors) Around two-thirds of our experimental participants make risk-neutral

(Insert Table 2 about here.)

choices; only 24% to 29% (5% to 11%) make risk-averse (risk-preferring) choices. Furthermore, in all of the experiments reported for the main results with parameter settings (π_w=.8,π_b= .1) and (.45, .1) over half of the participants in each session make risk-neutral choices. *Ceteris paribus,* this may tend to aid the experimental realisation of the theoretical predictions as CL assume risk neutral expected payoff maximisation.

In the experiments reported below we focus primarily on two sets of initial beliefs; (.45, .1) for the *benign* case and (.8, .1) for the *patronising* case. In particular we inquire if the following propositions can be verified experimentally:

Proposition 1. *Benign AA Equilibrium.* With initial starting point: $\pi_b < \pi_w < 0.5$, upon implementation of AA the proportion of *B*'s investing will rise so that $\pi_b = \pi_w$.

Proposition 2. *Patronising Equilibrium*. With initial starting point: $\pi_b < 0.5 < \pi_w$, upon implementation of AA the proportion of *B*'s investing will rise so that $\pi_b = 1 - \pi_w < \pi_w$.

Table 3 provides summary results of the Affirmative Action rounds of the experiments. Details follow but note the first row of Table 3 confirm proposition 1. In the three benign equilibrium sessions, π_b, closely tracks the *CL* predicted path in converging to the benign equilibrium, $\pi_b = \pi_w = .45$. Note in rows 2 and 3 proposition 2 is not confirmed for either (.8, .1) or (.6, .2). Rows 4 through 6 demonstrate that the prior results are robust to an alternate framing and to freeing up π_w to vary from round to round (based on prior W group experimental session results). Despite confirmation of the CL benign equilibrium, all experiments reported in Table 3 display substantial over-investment by experimental participants (relative to risk neutral behaviour).

(Place Table 3 about here.)

Figure 5 provides further details of row 1 Table 3 results for AA. The top panel summarising three sessions with ten participants each, shows that the employer beliefs about the proportion of *B*'s qualified for T_1 rise rapidly to .45 and, for the most part, stay

[Place Figure 5 about here.]

there, with just a little weakening in rounds 10 to 14. The *benign* equilibrium is reached and largely sustained, consistent with the CL theoretical prediction.

The middle panel in Figure 5 shows that over-investment is again observed. In this session, the over-investment did weaken a bit in rounds 8 through 13 where the actual path and the predicted path actually overlap for a couple of periods but after round 13 the over-investment reasserts itself and we observe investment rates climbing above the predicted path over the last two rounds.

Finally, the bottom panel in Figure 5 shows that under affirmative action the proportion of B's assigned to T_1 is equal to or exceeds the assumed proportion of W's. So

in terms of the updated value of π_b and the proportion of *B*'s assigned to T_1, these experimental sessions verify the *benign* equilibrium. Despite the over-investment, the system converges to what would look like a race-neutral outcome. *B*'s are investing at a high rate, a high proportion of *B's* are assigned to the more advanced task, T_1, and employers believe that *B*'s and *W*'s are equally likely to be qualified for T_1.[12] Thus an AA intervention is successful in eliminating the negative stereotype and leads to equal proportional representation in T_1 by race.

In the second set of experimental sessions, see Table 2 row 2, initial assumed beliefs are (0.8, 0.1). Theoretically this represents a stable negative stereotyping equilibrium of the patronising type. Implementing AA should theoretically lead π_b to converge to (1- π_w), or 0.2 in this particular case. We run three sets of experiments, with ten participants in each and again report the average results across the three experiments.

In the first round of AA where we start at (0.8, 0.1), the top panel of Figure 6 shows that the system does not converge to the predicted patronising equilibrium. We are typically left in middle ground, somewhat below the *benign* AA equilibrium where both

[Place Figure 6 about here]

beliefs are at 0.8 but clearly above the predicted patronising equilibrium where $\pi_b = 0.2 < \pi_w = 0.8$. Thus unlike the (0.45, 0.1) case, this experimental session does not appear to converge to the equilibrium predicted by CL under affirmative action. It is worth noting, however, that the beliefs do appear to fit <u>*a*</u> patronising equilibrium, just <u>*not*</u> the one predicted by CL. The belief, π_b, rises rapidly to rough equality with π_w, but then after round 10 drops down to the .4 to .6 range where it remains. Firms continue to believe that *B*'s are less qualified than *W*'s and hence continue to patronise *B*'s.

Figure 6's middle panel verifies that there is again over-investment. It is this over-investment that keeps us from realising fully the predicted patronising equilibrium. Nevertheless by the time the system settles down, *B*'s are investing much less than *W*'s, but there is just not as great a separation as predicted by the model. As the bottom panel of Figure 6 verifies, the affirmative action requirement is satisfied as *B*'s and W's are assigned to T_1 in essentially the same proportion.[13]

From a policy perspective AA may be expected to lead B workers to invest less-due to patronage i.e. an increased probability of being assigned to T_1 without investing. Interestingly at least in an experimental setting under-investment due to the patronage option is not observed; over-investment once again prevails. In the end we are neither in the *benign* equilibrium nor the patronising equilibrium predicted for this set of parameter values. We are in a patronising affirmative action equilibrium where affirmative action requirements are being met, about half as many *B*'s are investing compared to the *W* rate, employers believe *B*'s less likely to be qualified compared to *W*s, and there is still substantial patronage, with 50 percent or more of failing *B*'s being assigned to T_1. Thus the CL theoretical prediction is not perfectly verified for this patronising case.

Extensions and sensitivity

We run three additional experiments with slightly different treatments to examine the robustness of the above results. Given the failure to arrive at the CL predicted patronising equilibrium in the (.8, .1) case, we select an alternative set of 'patronising' beliefs ($\pi_w = 0.6$, $\pi_b = 0.2$). We run four sessions with ten participants in each session.[14] Because the initial three sessions were done as pilots and have fewer rounds, we only briefly discuss the one set of results with a full ten pre-AA and 15 AA rounds. The results

of the other sessions are, however, perfectly consistent with the reported results. Once again there is substantial over-investment and consistent with earlier results, we once again fail to reach the predicted patronising equilibrium when AA is introduced (See Figure 7.) In fact, in all four sets, the experiments now converge instead to the benign equilibrium, $\pi_b = 0.6$. These results suggest that the outcome is somewhat sensitive to

[Place Figure 7 about here.]

parameter values within the set of beliefs that are predicted to lead to a patronising equilibrium. It may be that an increase in the initial π_w or an increase in the spread between π_w and π_b make a patronising equilibrium outcome more likely

Is the tendency to over-invest influenced by framing the outcome of investing as being the opportunity to pass or fail a test?[15] To examine this notion, we introduce a version of the experiment where the terms, 'pass' and 'fail' are not used at all. In this alternate (neutral language) scenario, participants are given the choice to enrol in an online accreditation programme that may or may not be recognised by employers. There is no question of passing or failing a test. We find this re-framing has no impact on over-investment. In both sessions (not shown) where the alternate language is employed, B workers over-invested substantially as before. The confirmation (non-confirmation) of Proposition 1 (Proposition 2) is robust to the framing change.

We consider one final variation of the experiment. We allow firm beliefs about the qualification of W's, π_w, to evolve. As over-investment by B's appears to be endemic, we allow for that possibility among W's. The W's face a certainty effect in pre-AA and AA rounds. Since the evolution of π_b has no impact on the decisions of W's, we run a separate W experiment for one of the patronising cases, ($\pi_w = 0.6$, $\pi_b = 0.2$).[16] Then, in a

separate experiment for B's, rather than facing a fixed π_w every period, π_w is varied from round-to-round to match the pattern observed in the W experiment. Figure 8 shows that, in ten of the sixteen rounds, the value of π_w is at 0.8 or above. This high value for π_w has two effects on B's under the AA treatment. First, recall that α, the rate of patronisation, rises with π_w and the incentive to invest falls, somewhat offsetting the tendency to over-

[Place Figure 8 about here.]

invest and arrive at the *benign* outcome. Second, the evolution of the value of π_b may be capped at a higher level. In principle this should give more room for π_b to rise without bumping up against π_w. That, too, should make the *benign* outcome less likely. Nevertheless we observe in figure 8 that the value of π_b chases π_w up to high levels, finally matching the higher series in the 13th period and staying even with it to the end at the 16th period, even after π_w rose one last time. We arrive at the benign equilibrium, as before. Hence our main results are robust to all three of these variations. Our results tend to confirm Proposition 1 but not Proposition 2.

Further Discussion:

All of our experimental sessions led to higher investment rates than predicted by CL. In the benign case, the over-investment did not prevent experimental attainment of the predicted equilibrium. However, in the patronising case, the effect appears to diminish the likelihood of arriving at the CL predicted patronising *AA* equilibrium.

Recall our risk profile analysis suggests that the majority of experimental subjects are risk neutral. So, the observed substantial over-investment (relative to risk neutral behaviour) in both pre-AA and AA experimental rounds is a puzzle. The certainty effect of Kahneman and Tversky (1979) may provide some insight. Under affirmative action, if

the worker invests, paying a cost, *c*, a payoff of 1 is certain. If the worker chooses not to invest, the worker receives a payoff of 1 with probability $1 + (\alpha - 1)\pi_w$ and zero with probability $(1 - \alpha)\pi_w$, where $\alpha = \frac{\pi_w - \pi_b}{1 - \pi_b}$. In the patronising case, (.8, .1), the worker faces a choice between $1 - c$ for certain (if investing) and a lottery between a payoff of 1 with probability .823 and a payoff of 0 with probability .177. An expected value maximiser will choose to invest only when $c < .177$. As *c* is in the unit interval and is randomly assigned, only 17.7% of *B*'s should invest under Affirmative Action. We find much higher rates of investment (40 to 50% or even higher), consistent with the over-weighting of certain outcomes relative to risky ones with higher expected values as identified by Kahneman and Tversky.[17]

In the patronising case the problem of excess investment is exacerbated by the belief updating mechanism under Affirmative Action. Recall that in AA all B's who invest are assigned to T_1. Hence the firm accurately updates its belief, π_b, to the correct proportion. Over-investment will, therefore, lead to a rise in π_b which in turn leads to a reduction in α. A lower level of patronage implies a greater incentive to invest and thus greater divergence from the CL predicted adjustment path.

Because over-investment, relative to that anticipated for risk-neutral expected payoff maximisers, hinders the attainment of the CL patronising equilibrium, a fruitful way forward for future research might be to rework the CL model using a non-expected utility approach. An alternative way forward, suggested by the work of Carlin (1990), would be to search for a display or frame for presenting the worker investment lotteries that might aid the participants in their choices.

Finally, what do we learn about the efficacy of our constrained belief updating mechanism compared to the full belief updating mechanism in CL? Pre-AA results in Figure 9, for (.45, .1) show that the negative stereotype would be maintained given our constrained updating, but is volatile with CL updating, rising well above the 0.1 prediction. Once again substantial over-investment is observed but a clear separation between *B*'s and *W*'s theoretically predicted investment levels prevail. Finally despite substantial over-investment there is still a large separation between the proportion of *B*'s and the proportion of *W*'s who are assigned to T_1.

IV. Conclusions

The experimental analysis has successfully replicated many of the CL theoretical predictions. In particular we are able to demonstrate the potential 'stickiness' of a negative stereotype, and to replicate the impact of the introduction of AA in the benign case. However, the experiments suggest widespread over-investment relative to what one would expect by risk neutral optimising agents. A plausible explanation of this is the certainty effect, where participants prefer the certainty of a positive outcome associated with investing relative to an uncertain payoff with a higher expected value. The certainty effect, or other non-expected utility choice behaviour, leads to over-investment which tends to disrupt the theoretically predicted equilibrium in the context of the patronising case. The CL benign equilibrium is supported but not (the precise) patronising equilibrium.

Implementing the type of AA program considered here leads to an increase in the probability of a non-investing (i.e. unqualified) B worker being assigned to Task 1. That raises the policy concern that B workers will rely on patronage, choosing not to invest,

thus reinforcing the negative stereotype. This concern is *not* supported by the experimental analysis. At the same time that AA makes it possible for B's to be assigned to T_1 without investing, it also makes it *certain* that B's who do invest will be assigned to T_1. Once the proportion of investors exceeds the theoretical expectation, dynamic adjustment reduces the likelihood that beliefs fall back to the predicted level. These results indicate that in many potentially *patronising* situations, the certainty effect *may* help to dissipate patronage. Future work with non-expected utility maximising models or alternative displays of choice lotteries may shed further light on this question.

Figure 1: CL General Case—Negative stereotyping equilibria

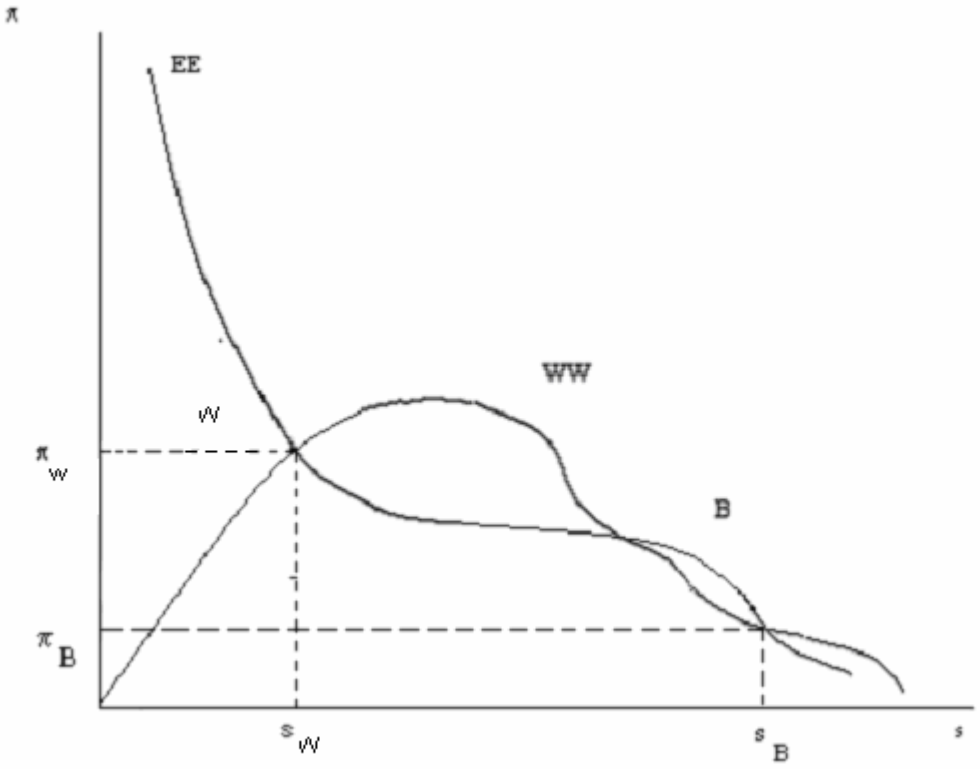


Figure 2: Employer-determination of standard

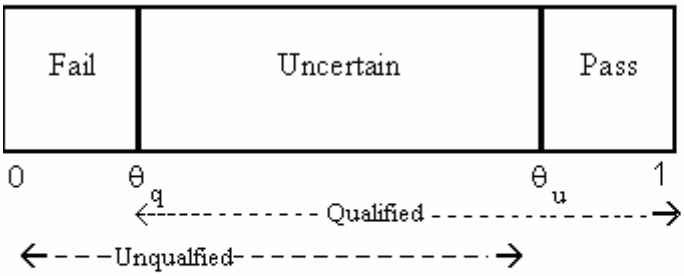


Figure 3: Investment, test score and task allocation

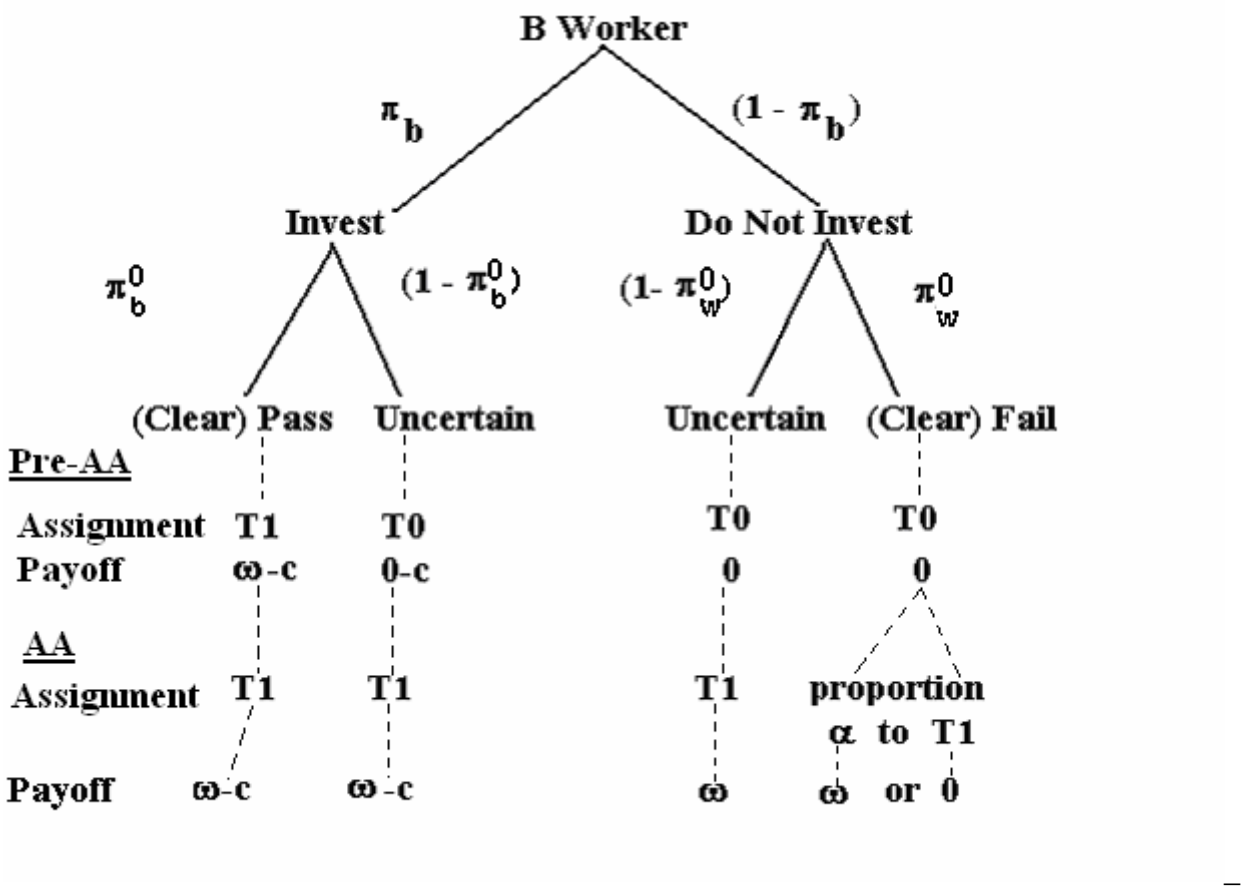


Figure 4: Flow Chart for Basic Stereotyping Experiment (without AA treatment)

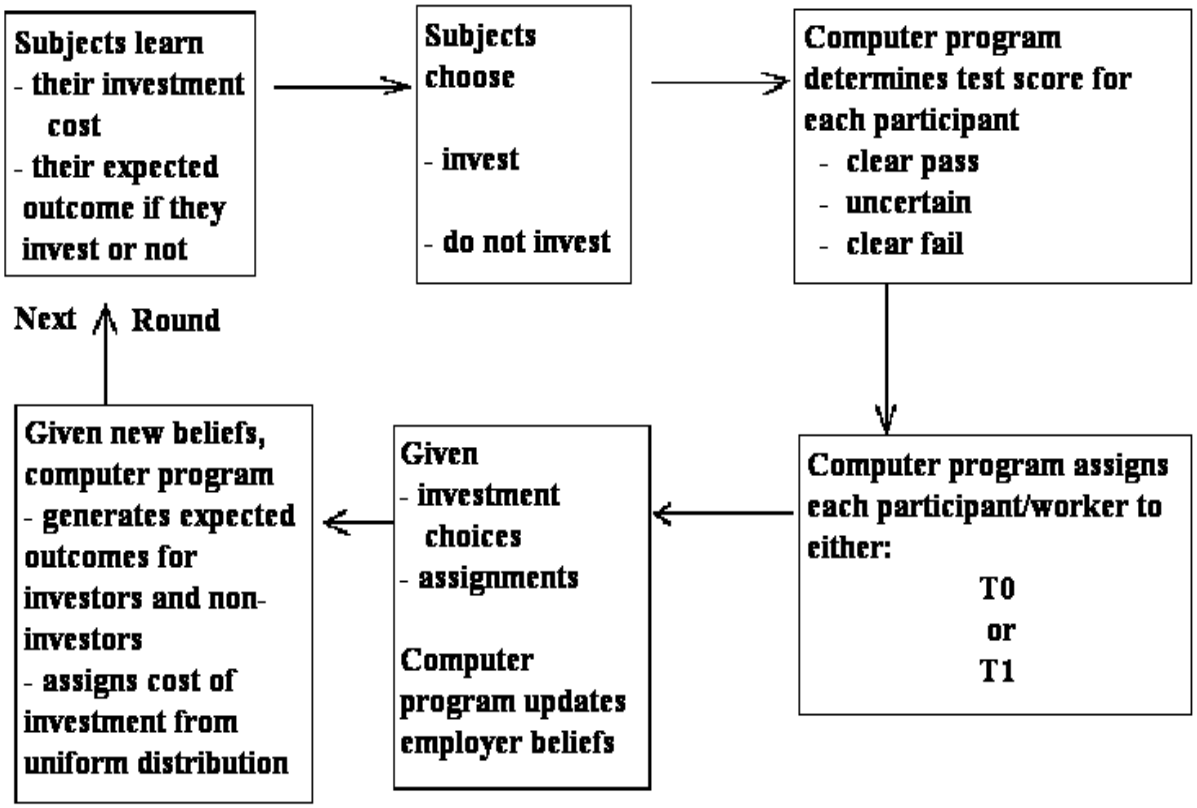


Table 1: Basic Structure of Experimental Design

Session Number	**Date**	**Time**	**Place**	**Treatment (π_w, π_b)**	**Framing**	π_w
1. benign	5/23/06	10:12 am	Deakin	(.45, .1)	Primary	Fixed
2. benign	5/23/06	12:04 pm	Deakin	(.45, .1)	Primary	Fixed
3. benign	5/23/06	1:06 pm	Deakin	(.45, .1)	Primary	Fixed
4. patronising	5/12/05	2:12 pm	Curtin	(.8, .1)	Primary	Fixed
5. patronising	5/31/05	2:48 pm	Curtin	(.8, .1)	Primary	Fixed
6. patronising	5/26/06	12:03 pm	Deakin	(.8, .1)	Primary	Fixed
7. patronising	5/29/06	12:06 pm	Deakin	(.6, .2)	Primary	Fixed
8. patronising	5/31/05	5:28 pm	Curtin	(.6, .2)	Primary	Varying
9. benign	5/18/06	12:24 pm	Deakin	(.45, .1)	Alternate	Fixed
10. benign	5/22/06	12:12 pm	Deakin	(.45, .1)	Alternate	Fixed
11. patronising	5/31/06	10:12 pm	Deakin	(.8, .1)	Alternate	Fixed

All experimental sessions included 10 subjects with 10 pre-Affirmative Action (pre-AA) rounds, following 2 practice rounds, and 15 Affirmative Action (AA) rounds, following 1 additional practice round.

Table 2: Prior Questions to Elicit Risk Profile

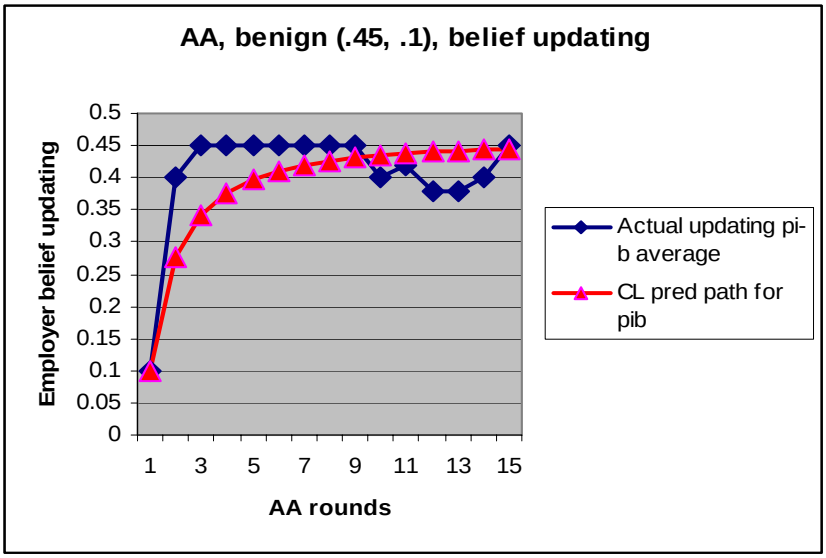
Year	Percentage of experimental subjects in each category		
	Risk Averse (AA)	**Risk Neutral (AB)**	**Risk Preferring (BB)**
2005	**24**	**65**	**11**
2006	**29**	**66**	**5**

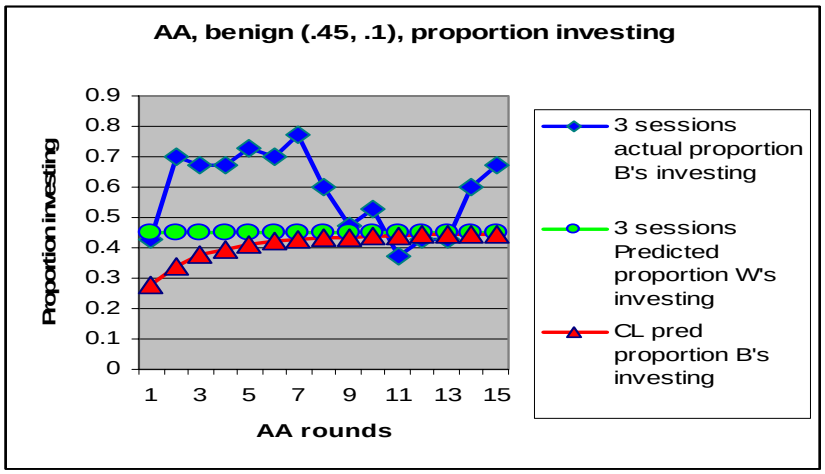
Table 3: Characterisation of Affirmative Action Results

Session Numbers	**Treatment (π_w, π_b)**	**Variation***	**Characterisation of Findings**
1, 2, 3	(.45, .1)	Standard	π_b closely tracks its *CL* predicted path to converge to π_w = 0.45; substantial over-investment. Proposition 1 confirmed.
4, 5, 6	(.8, .1)	Standard	Initial convergence of π_b path to π_w but after 8th round, π_b tracked down and then stayed in .4 to .55 range, above the *CL* prediction of 0.2; substantial over-investment. Proposition 2 not confirmed.
7	(.6, .2)	Standard	π_b converges to π_w = 0.6, rather than to *CL* predicted value of 0.4; substantial over-investment. Proposition 2 not confirmed.
8	(.6, .2)	Varying π_w	π_b converges to π_w = 0.6, rather than to *CL* predicted value of 0.4; substantial over-investment. Proposition 2 not confirmed.
9, 10	(.45, .1)	Alternate Framing	π_b closely tracks its *CL* predicted path to converge to π_w = 0.45; substantial over-investment. Proposition 1 confirmed.
11	(.8, .1)	Alternate Framing	π_b converges to range (0.2, 0.5) rather than to *CL* prediction, 0.2. Proposition 2 not confirmed.

***** The phrase 'Varying π_w' means that π_w is varied from round to round according to prior experimental results; 'Alternate Framing' refers to language describing the choices in a way that does not make use of a testing scenario which may have, but apparently did not, influence the results. 'Standard' means that the primary framing indicated in the Appendix is used and that the values for π_w are fixed at initial levels throughout the session.

Figure 5: AA, *benign* case, (0.45, 0.1), 3 sessions of 10 participants each– evolution of firm beliefs about likely qualification of *B*'s, the proportion of *B*'s and *W*'s investing, and the proportion assigned to T_1





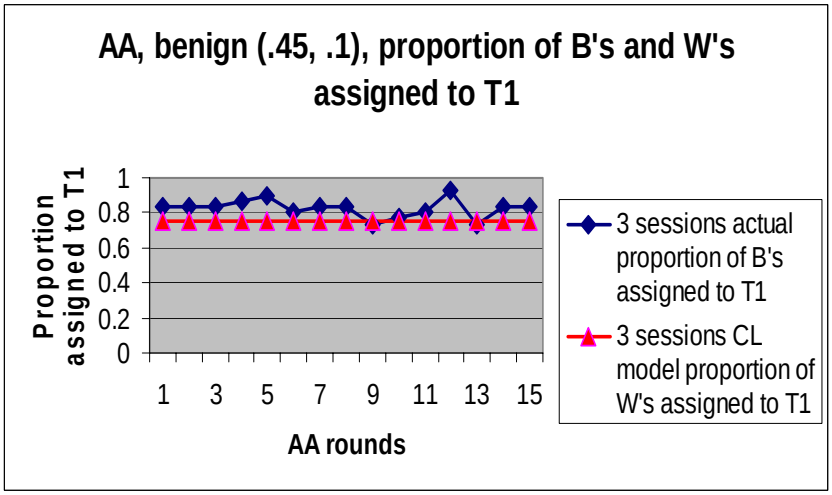


Figure 6: AA, *patronising* case, (0.8, 0.1), 3 sessions of 10 participants each– evolution of firm beliefs about likely qualification of *B*'s.

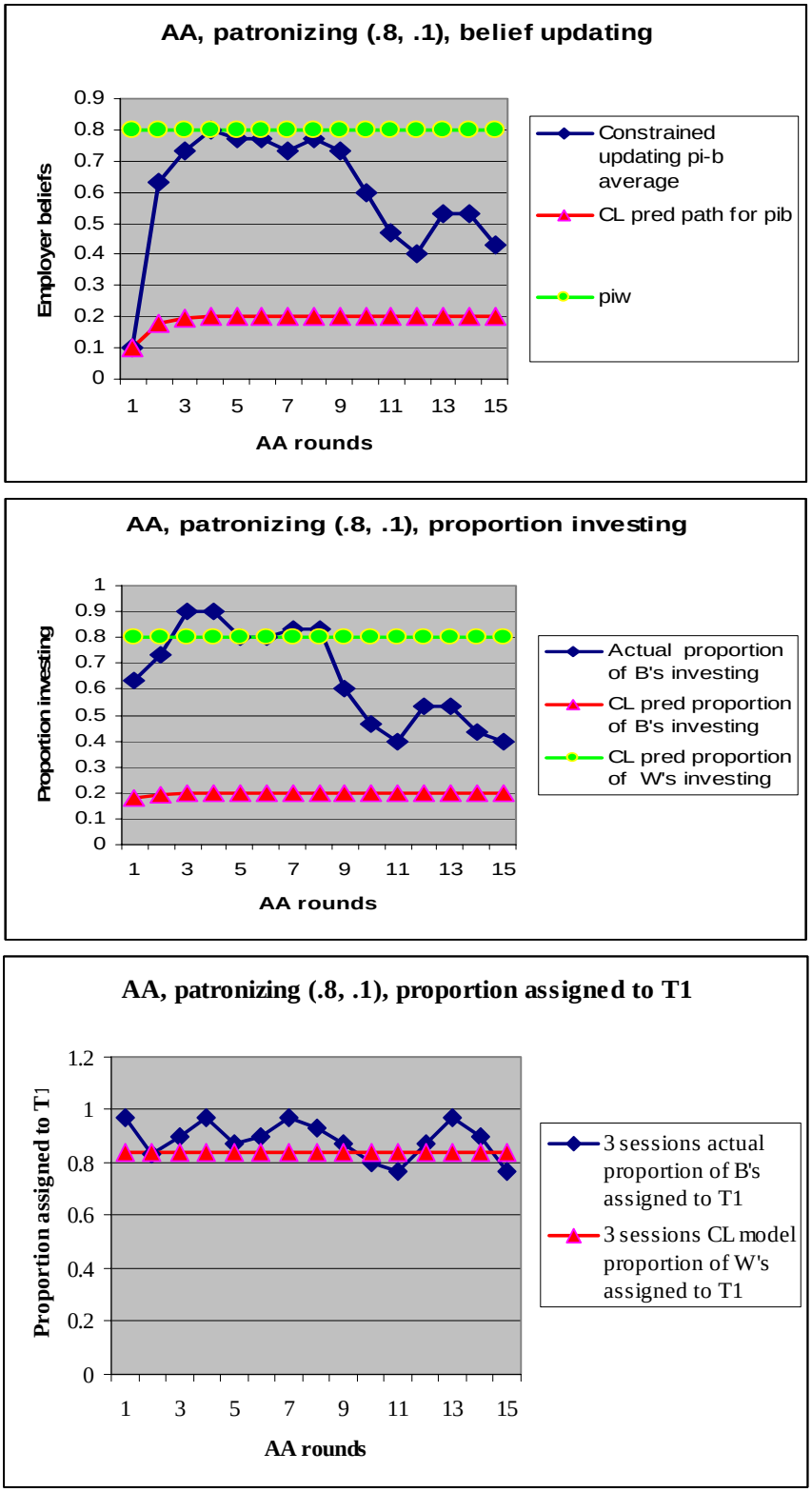
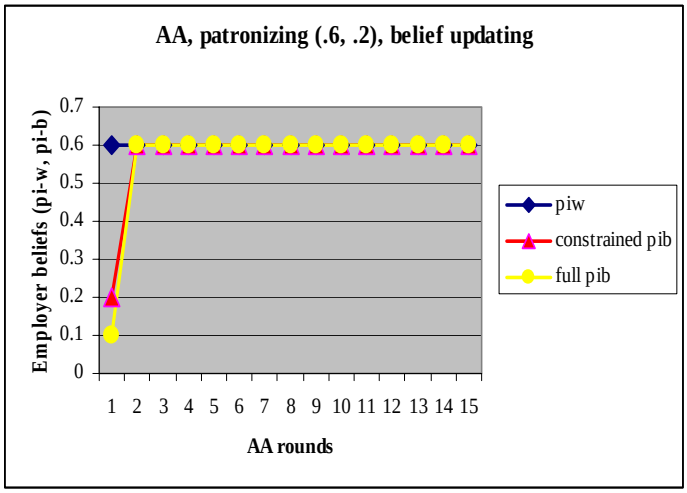
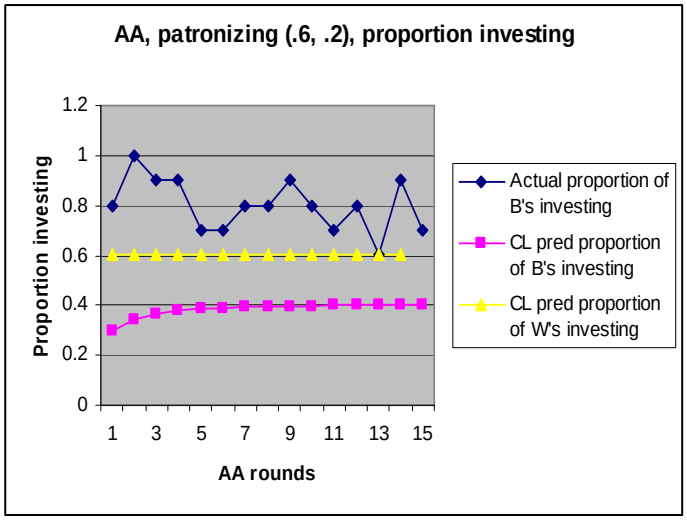


Figure 7: AA, *patronising* case, (0.6, 0.2), 1 session of 10 participants– belief updating, investment and actual proportion of *B*'s assigned to T_1 compared to assumed proportion or *W*s assigned to T_1.





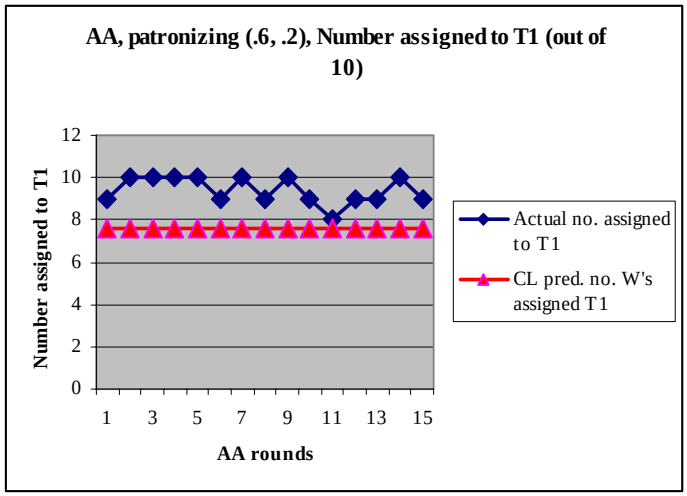


Figure 8: Evolution of beliefs, π_b, in an experiment with *B*'s when π_w varies according to prior experimental results (0.6, 0.2) [when participants were *W*'s]

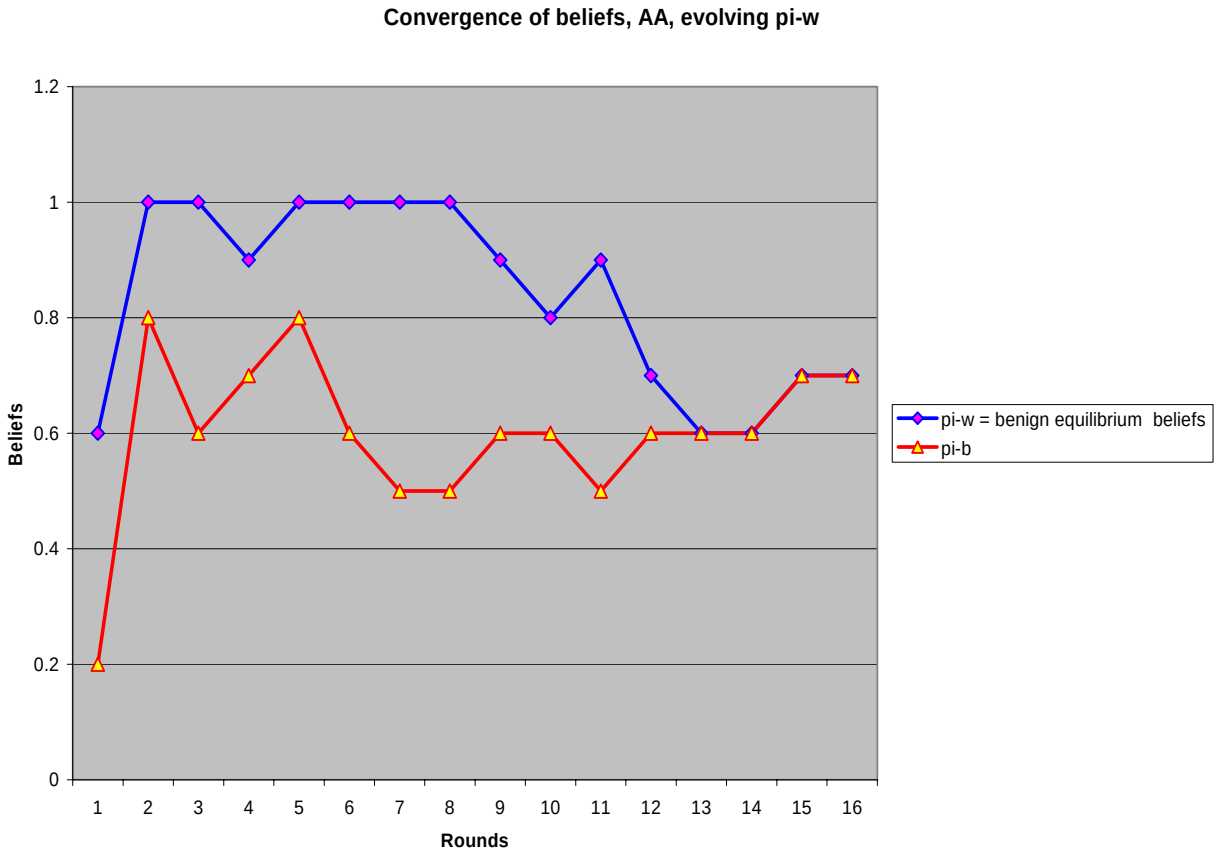
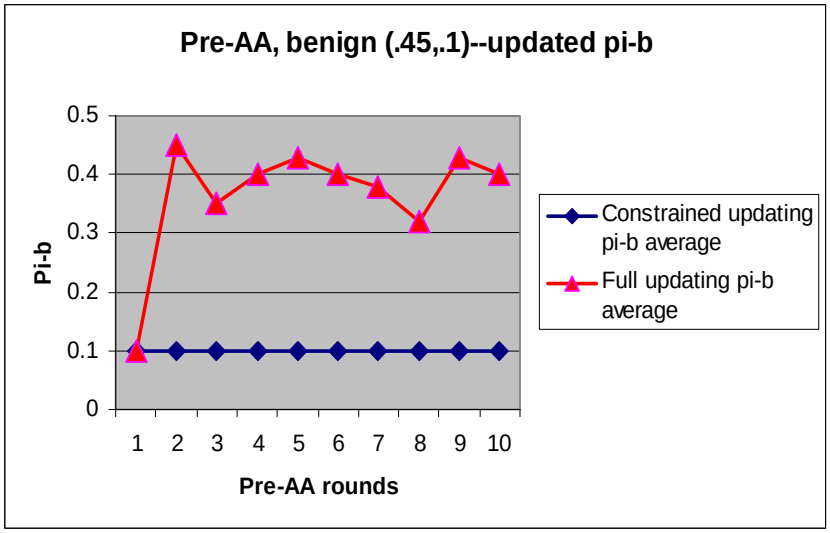
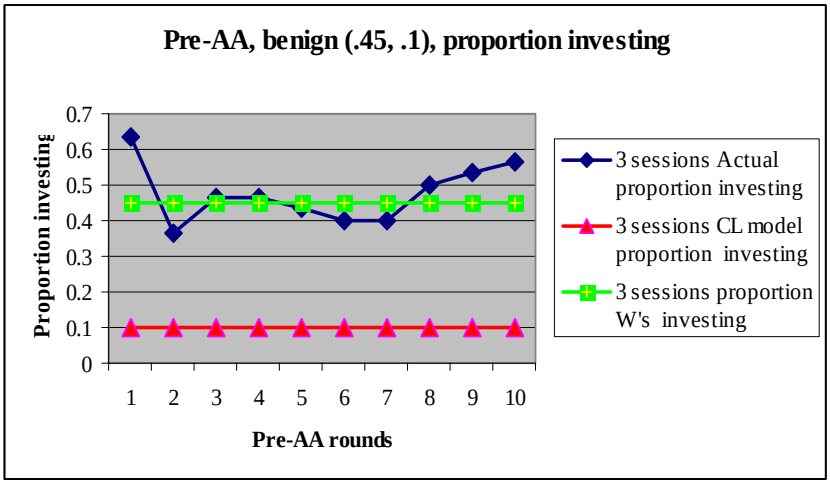
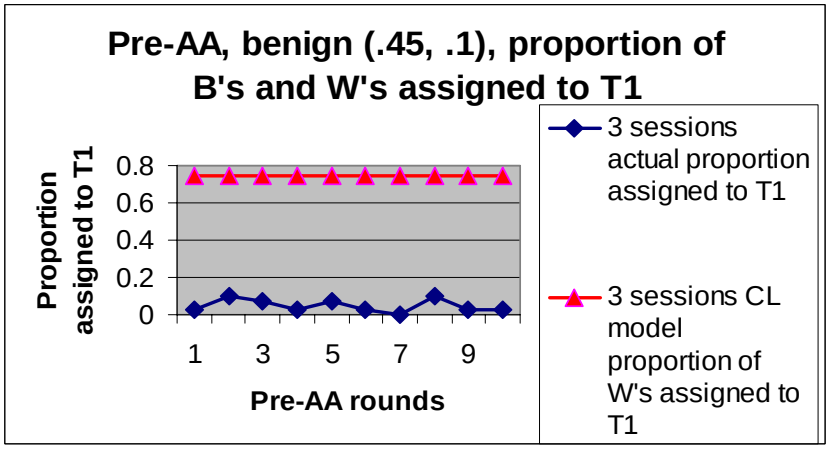


Figure 9: Pre-AA, *benign* case, (0.45, 0.1), 3 sessions of 10 participants each – evolution of firm beliefs about likely qualification of *B*'s, investment proportion and proportion assigned to T_1







REFERENCES

Anderson, Lisa, Fryer, Roland and Holt, Charles A. “Discrimination: Evidence from Psychology and Economics Experiments.” in W. Rogers, ed., *Handbook of Discrimination*, (forthcoming).

Calegari, Michael J., Schatzberg, Jeffrey W., and Sevcik, Galen R. “Experimental Evidence of Differential Auditor Pricing and Reporting Strategies.” *The Accounting Review*, 1998, 73 (2), pp. 255-275.

Camerer, Colin. “Individual Decision Making.” In J.H. Kagel and A.E. Roth, eds. *The Handbook of Experimental Economics*, Princeton University Press: Princeton, New Jersey, 1995, pp. 587-703.

Carlin, Paul S. “Is the Allais Paradox Robust to a Seemingly Trivial Change of Frame?” *Economics Letters*, 1990, 34, pp. 241-244.

Cason, Timothy N. and Friedman, Daniel. “Customer Search and Market Power: Some Laboratory Evidence.” June 1999. Unpublished Manuscript.

Cason, Timothy N. and Sharma, Tridib. “Recommended Play and Correlated Equilibria: An Experimental Study” August 12, 2006. Forthcoming in *Economic Theory*.

Coate, Stephen and Loury, Glenn C. "Will Affirmative-Action Policies Eliminate Negative Stereotypes?" *American Economic Review*, 1993, 83 (5), pp. 1220-1240.

Cox, James, Jason Sachat and Mark Walker. "An Experiment to Evaluate Bayesian Learning of Nash Equilibrium Play." *Games and Economic Behavior*, January 2001, 34 (1):11-33.

Datta, Shakun. "An Experimental Study of the Negative Externality from Improved Information in a Differentiated Products Search Model." June 2005. Unpublished Manuscript.

Feltovich, N. and Papageorgiou, C. "An experimental study of statistical discrimination by employers" Southern Economic Journal 70 (2004), pp 837-49.

Fryer, Jr., Roland G., Goeree, Jacob K. and Holt, Charles A. "CLASSROOM GAMES: Experience-Based Discrimination." *Journal of Economic Education*, 2005, 36 (2), pp. 160-170.

Holt, C.A. Markets, Games, & Strategic Behavior, Pearson Education Inc. (2007);

Holt, Charles A. and Laury, Susan K. "Risk Aversion and Incentive Effects." *American Economic Review*, 2002, 92 (5), pp. 1644-1655.

Kachelmeier, S.J. and Shehata, M. "Examining Risk Preferences under High Monetary Incentives: Experimental Evidence from the People's Republic of China." *American Economic Review*, 1992, 82, pp. 1120-1141.

Kahneman, Daniel and Tversky, Amos. "Prospect Theory: An Analysis of Decision Under Risk." *Econometrica*, 1979, 47 (2), pp. 263-292.

Kübler, Dorothea, Müller, Wieland and Normann, Hans-Theo. "Job Market Signaling and Screening: An Experimental Comparison," unpublished working paper.

Loomes, Graham and Sugden, Robert. "Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty." *The Economic Journal* 1982, 92 (368), pp. 805-824.

Spence, A. M. "Job Market Signaling." *Quarterly Journal of Economics*, 1973, 87(3), pp. 355-374.

Appendix

I. General Instructions

Primary Framing:

Object. We are trying to understand how individuals make investment decisions in the presence of risk about the rate of return. In this experiment you will be given a virtual **card** that has the specific scenario that you must use to determine your own investment choice.

In brief. The scenario is as follows: there are two types of tasks: Task 0 and Task 1, where Task 1 pays more than Task 0 (more specific details will be provided on your **card**). You can influence your chance of being allocated to the higher paying task (Task 1) by **investing** in skills. The hiring firm uses a **test** to determine whether you are allocated to Task 0 or Task 1. If you invest you will receive either a **clear pass** or an **uncertain** on the test. If you do not invest you will receive either an **uncertain** or a **clear fail**. You will be told on your virtual card the rule the firm uses to allocate workers between Task 1 and Task 0. Investing in skills improves the probability of being assigned to Task 1.

Your decision. In each case your only decision is **whether or not to invest** in acquiring skills which will help you prepare for the test - which will in turn increase your chances of being assigned to Task 1.

Consequences of your investment decision (Job Assignments):

In each round you will be assigned to either Task 0 or Task 1 based on your test result:

- Task 0 - pays 0 points
- Task 1 - pays 100 points

Everyone's **card** is different. Please read your **card** very carefully - it will explain your own particular scenario in detail. You must decide **whether or not to invest** in acquiring skills to prepare for the test.

Cost of investing. Each **card** will also contain your cost of investing in the test preparation. These costs will range between 0 and 100 points.

Payment and Investing in skills:

In addition to the fixed participation payment (to be announced), you can earn/lose money in each round of the experiment.

- In each round in which you decide to not invest in acquiring skills for the test, you will receive the payment for the Task to which you are assigned.
- In each round in which you decide to invest in acquiring skills for the test, the cost of investing will be deducted from the payment for the Task to which you are assigned.

So there are four possible outcomes in each round of the experiment:

1. If you invest and you are assigned to Task 1 you will receive a net payment of 100 points minus your cost of investing.
2. If you invest and you are assigned to Task 0 you will receive a net payment that is negative; i.e. the cost of investing.
3. If you do not invest in the preparation/skills and you are assigned to Task 1 you will receive a net payment of 100 points.
4. If you do not invest in the preparation/skills and you are assigned to Task 0 you will receive a net payment of 0.

Remember, however, that those who invest in preparation tend to score higher on the test, and those who score higher on the test are more likely to be assigned to Task 1.

Overall Payment. You will go through this exercise several times. Each time you will receive points depending upon the Task to which you are assigned and your cost of investing (if you invested). Your points will be recorded.

Your total payment will be calculated as follows: The computer will randomly select 3 rounds from those in which you participated. The net payments for the 3 rounds will be summed and divided by one hundred to convert to dollars. This sum may be positive or negative. This is turn will be multiplied by a factor (to be announced) and added to your initial participation fee. Finally this sum will be rounded to the nearest 10 cents.

Private Information. All of the information about your costs of investing, probability of being assigned to Task 0 or 1, your actual assignment, and your gross and net payoffs is

private information. **Please do not discuss this information nor disclose it to any of the other participants in the experiment.**

Rounds. Between each round and the next, the calculations for the values for the **cards** for the next round will be calculated by the program. Once the calculations are complete, each participant will receive a new **card**. Please read the information carefully and make your choice of whether or not to invest for that round.

II. Questions Eliciting Subject Attitude Toward Risk

Having completed the background information the participants were asked to also consider the following question relating to risk, derived from Holt and Laury (2002).

Before we begin the experiment we would like you to focus on the following question. Please note that we are interested in finding out what you would prefer. This is not a test question—there are no wrong answers.

Two different scenarios (lotteries) are listed below. In each case please choose the lotteries you prefer (circle A or B).

For example, in the first choice,
Lottery A offers a 3 in 10 chance of winning $2.00 and a 7 in 10 chance of winning $1.60; Lottery B offers a 3 in 10 chance of winning $3.85 and a 7 in 10 chance of winning $0.10. Please circle the Lottery that you would prefer.

Circle Your Choice		**Lottery A**	**Lottery B**
A B	(3)	3/10 $2, 7/10 $1.60	3/10 $3.85, 7/10 $0.10
A B	(6)	6/10 $2, 4/10 $1.60	6/10 $3.85, 4/10 $0.10

III. Summary History: Participants could access round history at any point during the experiment. This provides a summary of their own decisions together with allocations to Task 1/0 and associated net payouts.

Round Decision Task(0,1) Cost Net payout Cumulative Payout

1

2

3

4
---
5 ETC…

IV. Payment Scheme

The payments for participants were based on their performance in the experiment. Let x represent the participant's net return in a given round of the experiment. In the Coate and Loury model as presented in the text, the payment would range from –1 to +1, but in the experiment, the net return, x, ranged, in principle, from –100 points to +100 points. Of course, no one who made optimal, risk neutral decisions should ever have a negative net return. In practice negative scores were quite rare.

Three rounds were randomly selected at the end of play and the points from these three rounds were summed, $S = \sum_{i=1}^{3} x_i$.

A weight was applied to S to determine the actual payment participants received. The points were divided by 100 and multiplied by a weight of 2.

V. Project Website. There is a project website that allows anyone to run experiments like the one reported in this paper. It is best to obtain a Supplementary Appendix (available from authors, anonymous for now) which has detailed instructions on how to run the experiment on the website as administrator.

Notes

[1] In fact CL suggest that AA may under certain conditions accentuate the negative stereotype.

[2] This paper has an accompanying web site: http://lab.ohwhoknows.com/admin. The Login name is guest and password welcome. There is on-line help providing complete instructions for running your own classroom experiments. The computer program/software is freely available for both teaching/research purposes.

[3] The likely reason for the separation is random cost differences; in this case purple workers drew higher costs than green workers in early rounds.

[4] This is in contrast to the FP one-sided experiment in which experimental participants play the role of firms. A companion experiment reversing the roles is under development.

[5] Future work hopes to use the insights provided by analysis of worker behaviour in the current paper and firm behaviour in the mirror image of the current analysis to shed additional light on the full interaction examined by FGH.

[6] As long as the absolute value of the slope of EE exceeds the absolute value of the slope of WW at s^* then CL show that the equilibrium is locally stable.

[7] It is argued that more subtle forms of AA (e.g. assigning all B and W workers who achieve a score above some cut-off to T_1) are difficult to implement due to issues of observability and enforcement.

---

[8] Please see the supplementary appendix (available upon request) for a numerical example.

[9] This focus on B's assumes that λ, the proportion of W's in the population, is sufficiently large that optimal firm behaviour in the presence of AA will always be to adjust on the B side rather than the W dimension. In particular the ratio, λ, must exceed $\hat{\lambda} = \frac{1}{\xi_w(1+r)}$, where $\xi_w = \frac{1}{1+\left[\left(1-\pi_w \big/ \pi_w\right)\left(\frac{p_u}{p_q}\right)\right]}$, and all other terms are as previously defined.

[10] If B workers over-invest but are not assigned to T_1, we assume the firm cannot observe their ability to perform T_1 and hence beliefs are unaltered. Only if the over-investment leads to over-assignment will the firm belief alter. Alternatively if B workers substantially under-invest, the proportion of failing test scores (PCF_B) may be higher than compatible with the initial π_b. Specifically if PCF_B rises above $(1 - \pi_b)$, firm beliefs about the proportion of B's who are qualified is revised from π_b to $1 - PCF_B$. Given only B's with a pass on the test are allocated to Task 1, some over and under-investment will not be directly observed by the firm and consequently beliefs will be unaltered.

[11] The deliberately sparse experimental environment means that we learn only about responses to the basic incentive structure of the CL model. We have nothing to say in this paper about questions relevant to sociological or psychological concerns about the effects of racism.

[12] The equivalence of employer beliefs about *B*'s and *W*'s is due, in part, to our decision to cap the updating of π_b at π_w. This seemed appropriate for pre-AA as $\pi_b > \pi_w$ would imply no negative stereotyping and no need for affirmative action; it seemed appropriate for AA as $\pi_b = \pi_w$ would achieve the desired affirmative action outcome.

[13] In some of these sessions, it appears that the oscillations are damping and π_b is approaching the predicted level. It raises the hope that if the sessions were extended, perhaps the system would settle into the predicted equilibrium. We tested that by extending the number of rounds for the (.8, .1) patronising case to 25 rounds in one session and 40 rounds in another. In both sessions, there were successive rounds where π_b fell toward the predicted level; in all cases, this was only temporary as π_b rose thereafter to levels above the CL prediction. We concluded that extending the sessions beyond 15 rounds was not productive.

[14] In an unreported set, we increased the number of participants from 10 to 20 with no apparent effect on the main findings.

[15] One other possibility might be that the result is related to the one-sided nature of our experiment. In a supplementary appendix we examine results from Holt (2007) and find evidence of over-investment in their two-sided experiment.

[16] π_b affects the probability of an investor attaining a clear pass rather than an uncertain score. This has no direct impact on the expected decision of W's as all W-investors are assigned to T_1 whether they receive a clear pass or an uncertain score.

[17] The certainty effect phenomenon is but one of many anomalous experimental outcomes that emerged as challenges to expected utility theory between 1970 and the early 1990's. [See Camerer (1995) for an excellent survey of the theories and experimental findings.]