ISSN 1936-5349 (print)
ISSN 1936-5357 (online)

HARVARD

JOHN M. OLIN CENTER FOR LAW, ECONOMICS, AND BUSINESS

AFFIRMATIVE ACTION AND STEREOTYPE THREAT

Anat Bracha
Alma Cohen
Lynn Conell-Price

Discussion Paper No. 805

1/2015

Harvard Law School
Cambridge, MA 02138

Affirmative Action and Stereotype Threat

Anat Bracha
The Federal Reserve
Bank of Boston

Alma Cohen
Harvard Law School,
Tel Aviv University
and the NBER

Lynn Conell-Price
Carnegie Mellon University

Abstract

This paper provides experimental evidence on the effect of affirmative action (AA). In particular, we investigate whether affirmative action has a ”stereotype threat effect” – that is, whether AA cues a negative stereotype that leads individuals to conform to the stereotype and adversely affects their performance. Stereotype threat has been shown in the literature to be potentially significant for individuals who identify strongly with the domain of the stereotype and who engage in complex stereotype-relevant tasks. We therefore explore this question in the context of gender-based AA for a complex math task. In this context, the stereotype is most relevant for women with high math ability, and the stereotype threat effects can be expected to work in the opposite direction to AA's competition effect that encourages women to compete. We find that, consistent with the presence of a stereotype threat, AA has an overall negative effect on the performance of high-ability women performing complex math tasks.

JEL Classifications: C91, I28, J16, J78, K19, K31

Keywords: Affirmative action, stereotype threat, gender differences, GRE performance.

We would like to thank Uri Gneezy, Muriel Niederle, Analia Schlosser, Tali Regev, and participants of the North American ESA Meetings in Santa Cruz and the 2014 SITE experimental economic workshop for helpful comments. The project was made possible by the generous financial support of the Foerder Institute for Economic Research and the Harvard Law School. This research has been conducted with IRB approval. The views expressed in this paper are those of the authors and do not necessarily represent the views of the Federal Reserve Bank of Boston or the Federal Reserve System.

1. Introduction

Affirmative action (AA) has long been a hotly debated subject. This debate can benefit from an improved understanding of the consequences of AA, particularly its effect on the performance of targeted individuals. This paper provides experimental evidence that contributes to obtaining such an understanding.

While women represent half of the general population, as well as half of all college-educated individuals, women are substantially underrepresented in many selective or high-level professional positions and occupations. For example, women are substantially under-represented among corporate directors and top executives; women constitute less than 20% both in the US and in EU countries, and their fraction of top executive positions in large public companies is below 5% in both the US and EU countries (Catalyst Census: Fortune 500 Women Board Directors 2013 and European Commission 2012). Similarly, the percentage of females among partners in US law firms is much smaller than their percentage among law firm associates (National Association for Law Placement 2013). As a further illustration, women in OECD countries are substantially under-represented among those holding jobs in science, technology, engineering, and math (STEM) (see Meeting of the OECD Council at Ministerial level 2011 and, for the US, see U.S. Department of Commerce, Economics and Statistics Administration 2011).

This underrepresentation of women in some professional groups and positions has led to substantial interest in measures aimed at increasing female representation in those capacities, as well as to the adoption or consideration of AA policies that provide women with an advantage, sometimes in the form of guaranteed participation. For example, in 2003 Norway became the first country to legislate gender representation on corporate boards, passing a law that required about 500 firms to raise the proportion of women on their boards to 40%. Such measures were subsequently adopted by other countries, including Belgium, France, Iceland, Italy, Malaysia, the Netherlands, and Spain. More recently, the European Commission has proposed legislation to ensure that, by 2020, 40% of non-executive directors are women (European Commission 2012).

While AA policies can lead to the increased selection of women taking performance as given, it is important to recognize that these policies can also, in at least two ways, influence the performance on which selection is based. First, AA policies can counter the tendency of women to avoid competing vigorously in mixed-sex situations (see, e.g., Gneezy et al. 2003, Niederle and Vesterlund 2007, Niederle and Vesterlund 2010, Sutter and Rützler 2010, Niederle et al. 2013, and the review by Croson and Gneezy 2009). This tendency (to avoid or not compete vigorously in competition) might lead women either to fail to win participation in certain professional groups and positions or to avoid the selection process altogether, taking alternative career routes instead. The countering of such tendencies by AA policies, which we label "the competition effect," has been identified experimentally by several important studies (Schotter and Weigelt 1992, Calsamiglia et al. 2013, Niederle et al. 2013, and Balafoutas and Sutter 2012). These studies, using a range of tasks, find that AA policies actually counter such tendencies and improve the performance of the intended beneficiaries.

However, there is evidence that reminding individuals of belonging to a group that stereotypically underperforms leads them to conform to this negative stereotype. For example, studies document that reminding African-American students of their race leads them to perform significantly worse on verbal Graduate Record Exams (GREs) (Steele and Aronson 1995) and that reminding women of their gender impairs their performance in math (Shih et al. 1999, Spencer et al. 1999).[1]

This paper focuses on the second potential effect of AA policies on performance, which involves the introduction of a stereotype threat. The stereotype threat effect operates through

[1] At the same time, Shih et al. (1999) show that reminding Asian women of their race leads them to perform significantly better. This study therefore demonstrates the more general concept of stereotype susceptibility whereby positively stereotyped groups can also be primed and will subsequently perform better. We also note that, although most studies in the stereotype threat literature find a stereotype threat effect, others find opposite effects (Wei 2009, 2012) or no effects (Stricker 1998, Fryer et al. 2008).

priming—that is, by reminding individuals of negative stereotypes relevant to them or by providing information about relevant differences in achievement across groups (see, e.g., Stricker 1998 and Spencer et al. 1999). AA, by definition, singles out disadvantaged or underperforming groups and therefore may unintentionally remind its beneficiaries of relevant negative stereotypes. We therefore hypothesize that the introduction of an AA plan can produce a stereotype threat, and we test this hypothesis below.

Stereotype threat effects typically occur when a task is complex and a negative stereotype is relevant to the task. Studies that have focused on testing for a stereotype threat have therefore used tasks such as the Scholastic Achievement Test (SAT) and the GRE (Steele and Aronson 1995 and Spencer et al. 1999; see Wei 2009 for a summary of lab results). Experimental studies of AA programs, however, have thus far focused on the competition effect and not examined contexts with real, complex, and stereotype-relevant tasks. For example, Schotter and Weigelt (1992) use an abstract task where (cost) disadvantage is randomly assigned; Calsamiglia et al. (2013) use performance on Sudoku puzzles, which is not related to an established stereotype. Similarly, Niederle et al. (2013) use a simple math task in which subjects sum five 2-digit numbers in a given amount of time, where the gender gap in math performance emerges only with more complex problem solving (Hyde et al. 1990).

While the tasks used in the above studies enabled researchers to test the competition effect, which was the focus of their studies, the tasks did not enable them to test for a stereotype threat effect. We therefore use tasks that facilitate such testing—in particular, the solving of complex GRE quantitative questions, an endeavor in which the average performance of women is both stereotypically and actually worse than that of men. To the best of our knowledge, our work is the first to provide experimental evidence that AA policies have stereotype threat effects.

As the literature suggests that the stereotype threat effect is stronger for the group that most identifies with the stereotype domain (e.g., Aronson et al. 1999, Steele et al. 2002), our analysis pays special attention to high-ability women. In our context, we hypothesize that women of high baseline ability in the GRE quantitative task would be especially troubled by the stereotype that women do not perform as well in math as men. We therefore conjecture and test whether the stereotype threat effect is especially significant for high-ability women.

To investigate whether and to what extent AA policies induce a stereotype threat effect, we conducted a lab experiment with a gender-based quota policy. Participants were assigned to groups of four (two men and two women) and asked to solve quantitative GRE questions; their performance was graded in accordance with the customary practice used in GRE and other standardized admission tests; that is, we rewarded each correct answer and penalized each incorrect answer. Participants competed for a monetary prize, which was awarded (i) to the top two performers regardless of gender or (ii) to the top two performers subject to a gender quota requiring at least one female winner. To further examine whether AA may serve as a prime generating stereotype threat, we varied whether participants in AA groups were presented with information on the superior average performance of men over women on the actual quantitative GRE exam (informational priming).[2]

We analyze the effect of gender quota on the performance of men and women by comparing their scores on the exam across the AA conditions, and we examine whether AA has a similar effect with or without informational priming. We find that, while the presence of an AA policy leaves men unaffected, it changes the performance of women. Women with low baseline ability perform significantly better in the presence of an AA policy and women with high baseline ability perform significantly worse. However, we find no evidence that results are driven by a single-sex competition effect.

Overall, we find a pattern that is consistent with the presence of both a competition effect and a stereotype threat effect. On one hand, there is a positive effect of AA, visible by the improved performance of low-ability women. On the other hand, there is a negative effect of AA on the performance of high-ability women. This is consistent with the stereotype threat effect as we find no evidence that high-ability women reduce their effort in response to the AA policy or that this effect is due to single-sex competition. The different overall effect of high- and

[2] Informational priming in this context refers to using objective information as a prime or stimulus of a negative stereotype.

low-ability women can be explained as the net effect of two factors: the greater incentive to compete due to the increased likelihood of winning the prize and the stereotype threat effect. The former is likely stronger for the low-ability women, while the latter is stronger for high-ability women. This is because the AA significantly increases the chances of low-ability women to win the prize while the high-ability women already have high chances of winning without the AA, and, at the same time, high-ability women are expected to identify more with the stereotype domain.

The remainder of our paper is organized as follows. Section 2 describes our experimental design and procedure. Section 3 describes and discusses our results. Section 4 concludes.

2. Experimental Design and Procedure

To test whether gender-based AA can trigger a stereotype threat effect, we used a between-subject experimental design with random assignment to gender-based AA in a competitive setting with incentivized performance. Participants answered questions from past quantitative GREs, a complex task used in actual graduate school admissions. Because men's average performance on such tests is known to be better than women's, we consider this setting appropriate for investigating the existence of a stereotype threat effect. To examine the effect of AA on performance, we calculated a score that penalizes guesses—as is standard practice on exams such as the GRE. Specifically, we awarded a point for each correct answer and subtracted a quarter point for each incorrect answer.

To check for ability-based heterogeneity in the response to the preferential policy, we designed the experiment in three rounds of 10-minute math exams. This design enables us to use the first round, with noncompetitive piece rate incentives based on the individual's score, as a proxy for ability.

In the second-round exam, which is the main focus of our analysis, subjects were randomly assigned to a group of two men and two women[3] and competed within this group for a bonus of $10 on top of their pay for performance. The groups competing in this round were randomly assigned to one of three conditions: the control condition (No AA), the AA condition (AA), or the AA and informational prime (AA-I) condition. In the No AA condition, the two subjects with the highest scores won the tournament and each received the $10 bonus. In the AA and AA-I conditions, the two subjects with the highest scores each received the $10 bonus, subject to a gender quota that required at least one woman to receive the bonus. That is, if the two highest scores were both earned by men, then the highest scorer and the highest female scorer earned the bonus. The AA-I condition was identical to the AA condition except that participants assigned to the AA-I condition also received an informational stereotype prime prior to the exam. We included this manipulation to compare the effect of the quota policy alone, which may convey information that acts as a stereotype prime, to the effect of a direct stereotype threat prime similar to primes used in previous studies (e.g., Spencer et al. 1999).

The direct prime was included in the description of the quota policy as follows: "*Since ETS statistics show that females quantitative GRE scores are consistently lower compared with males by about 15 percent, we set the following rule: The two participants with the highest score in the group of four (two men, two women) will get the bonus, as long as at least one of the two is a woman. That is, if neither of the participants with one of the top two scores is a woman, the bonus will be given to the participant with the highest score overall, and to the female participant with the highest score. In other words, one of the two winners must be a woman*."

Finally, in the third round, subjects were paid according to their scores, as in the first round. After completing the three rounds but before learning what they earned for them and whether they had won the bonus, subjects filled out a questionnaire in which they were asked to self-report their SAT scores (quantitative and verbal) and the extent to which they exerted effort on our exam. They were then informed of their earnings in the three rounds, told whether they had won the $10 prize, and paid privately in cash. Average earnings were $25.43.

[3] It was not possible for participants to identify the other members of their group.

The experiment was programmed using Authorware 7.01 and run on computers in the Harvard Decision Science Lab. In total, 248 subjects participated in the study—80 subjects in the control condition, 84 subjects in the AA, and 84 subjects in the AA-I condition—and each condition contained equal numbers of men and women. All subjects were undergraduate or graduate students from Harvard University recruited from the lab's subject pool, and their average age was 20 years. The self-reported average quantitative SAT score was 729.73 and the average verbal SAT score was 719.46.

3. Results

To investigate the effect of implementing gender-based AA policies on performance in a competitive math test and to explore whether AA acts as a stereotype threat prime for women, we focus on the effect of AA on the change in test scores between the second and first rounds. The test score is the appropriate measure to examine, as it determines subjects' payments and captures both the quantity and accuracy of their responses. We analyze the effect of AA on the second-round score using a weighted least squares (WLS) model that adjusts for a systematic relationship between variance in score and number of questions attempted.[4] To test whether the effect on high-ability women is different from the effect on low-ability women, we use the test score in the first round as a measure of ability and interact it with AA. We also explore the effect of AA on the number of questions attempted and answered accurately, measures that may capture changes in response strategy.

[4] A technical concern of using OLS is heteroskedasticity, where the variance in the second-round score may systematically increase with the number of questions attempted. This concern is simply due to the fact that with more questions attempted, the potential high and low scores are more extreme. On the basis of this relationship, we use a WLS model with weights proportional to the number of attempted questions in round 2. We get similar results whether we use the OLS model or bootstrap regression model.

3.1. The Net Effect of Affirmative Action

Table 1 presents descriptive information on gender differences in performance for each round. In both the baseline round and the second round, men scored higher on average than women. Specifically, in round 1, men's average score is 6.36 while women's average score is 5.65 (one-sided *t*-test yields p = 0.08), and in round 2, men's average score is 7.31 while women's average score is 6.45 (one-sided *t*-test yields p = 0.05). Examining the average number of questions attempted and the average accuracy, we find systematic gender differences that appear to reflect different response strategies: men answered significantly more questions in every round, whereas in two of the three rounds, women were slightly more accurate than men. However, the difference in accuracy is not statistically significant.

Given the gender difference in baseline ability and response strategy, we opt to analyze the effect of AA on the change in score between the first two rounds for each gender separately. In doing so, we control for ability using the first-round scores and for response strategy using the number of questions attempted. We also control for age and self-reported SAT scores (verbal and quantitative).[5]

First, we examine the effect of the AA policy by pooling all cases of AA together, whether or not informational priming was provided. We consider two sets of specifications: one with a dummy variable that equals 1 when the AA policy is in effect and zero otherwise, and the other that also includes an interaction term of the AA dummy variable with baseline ability, as measured by the first-round score. Table 2 reports the results of these two specifications. We find that, as expected and regardless of the specification considered, baseline ability (first-round score) has a negative effect on the change in scores between the first two rounds and is highly significant for women. This negative effect is expected since it is more difficult to improve when starting off from a higher baseline score. We also find that the self-reported quantitative SAT score has a positive significant effect on the change in scores for both men and women;

[5] There is no significant gender difference in average quantitative SAT score in our sample: women's average quantitative SAT score is 736 while men's is 718 (one-sided *t*-test yield p = 0.11).

however, the self-reported verbal SAT score does not. The number of questions attempted in the first round, which may capture response strategy,[6] is associated with a significantly higher change in scores for women (effect of 0.442 or 0.471, depending on the specification, both with p-value < 0.001). The effect for men is less than half that for women and is not significant.

Turning to examine the effect of AA on performance, we find that, regardless of ability (Table 2, column 1), AA has a positive albeit insignificant effect. When we allow for the possibility that high- and low-performing women are affected differently (Table 2, column 2), however, we find a positive and significant main effect of AA (1.889, p-value = 0.019), which declines with ability (–0.23; p-value = 0.035). This implies that the overall effect of AA is negative for high-ability women whose first-round score is over 8.21. Looking at percentiles, approximately 80% of women have a first-round score lower than 8 in round 1 which is very close to this cutoff; that is, AA has a positive effect on most women, but the projected effect of the policy is negative for the top 20%.[7] For men, there is an insignificant positive effect of AA regardless of ability.

To test whether the overall negative effect on the high-ability women is significant, and since sample size becomes an issue, we use the bootstrap method. For this exercise, we split the women in our sample into three groups according to their first round scores—low ability (score below 5), mid ability (score between 5 and 8), and high ability (score 8 or higher)—where the high ability group corresponds to the group whose overall effect of AA seems to be negative. We then calculate for each subgroup the difference in (mean) scores between round 2 and round

[6] Indeed, in a regression of the success rate in round 2 on the number of attempts, gender, and their interaction, we find that for a given number of attempts, women have significantly higher success rates (the main effect of gender is 0.168, p = 0.054), which diminish marginally with the number of attempts (the interaction is –.012 p = 0.107). Nevertheless, the gender effect is positive up to 14 attempts, representing about 92% of the women in our sample. This is consistent with a different response strategy across gender, where men attempt more questions at the cost of accuracy.

[7] Note that since we are working with a sample of students from a very selective university, the results for the lower range of the ability distribution in our sample may be more representative of a broader population than are the results for the entire sample. At the same time, AA policies often aim at the very best individuals within the beneficiary group, in which case the very best in our sample would be the most interesting and relevant individuals to examine.

1, we calculate it separately for women who were under AA and women who were in the control group, and we then take the difference-in-differences (AA minus control). We repeat this exercise 500 times, randomly sampling women in each subgroup with repetition, resulting in a distribution of difference-in-differences (see Figure 1). This allows us to ask whether the average diff-in-diff is negative and significant for the high-ability women subgroup. We find that the mean diff-in-diff for the high-ability women is negative (−1.35), slightly positive for the mid-ability women (0.156), and positive for the low-ability women (0.677). Looking at the distribution of the mean diff-in-diff results for the high-ability women shows that in 93% of the random samples, this value is found to be negative. For mid-ability women, the small positive effect is insignificant, and for low-ability women it is positive for 85% of the random samples.

The advantage of the bootstrapping method is that it doesn't require any parametric assumptions. However, we complement this exercise with a regression analysis by grouping women according to the three ability groups described above. The results we get (see Table 3) are consistent with those obtained from the bootstrap approach: the effect of AA on the low-ability women is positive (1.43) and significant (one-sided *t-test* yield p = 0.03); the effect of AA on mid-ability women is overall negative (−0.118) but insignificant, and the effect of AA on high-ability women is negative (−1.67) and significant (one-sided *t-test* yield p = 0.03)

The positive main effect of AA on women's scores reported in Tables 2 and 3 is not surprising given that the gender quota increases (at least weakly) the women's objective chance to win the bonus. The finding that the positive effect of AA decreases with ability—to the point that it has a negative effect on high-ability women—is surprising and suggests that there is another factor offsetting the main effect.

3.2. Does Affirmative Action Act as a Prime?

Can the observed negative effect of AA be due to AA acting as a negative prime? To address this question, we first examine whether the effect observed is driven by women exposed to the informational prime or could be obtained by AA alone. We also check whether the negative effect of AA can be explained by high-ability women reducing their effort and question whether reducing effort could be an optimal response.

Table 4 presents the results with a specification similar to Table 3 but also with an indicator variable "Info," which takes the value of 1 if a participant was assigned to the AA treatment with the additional informational prime.[8] We find that the Info indicator variable's main effect and its interaction with ability group are insignificant, suggesting that the effect of AA on women is the same whether or not participants receive a direct informational prime.

Using the results reported in Table 4 and the distribution of first-round scores for women, we find that women in the first two ability groups with first-round scores below 8—about 80% of the women in our sample—improved or did not change scores under the AA treatment. By contrast, women in the high-ability group—the top 20%—performed worse under AA. The results for men shown in Table 4 are very similar to those shown in Table 3, indicating that adding the informational prime dummy does not change the results.

Finding no effect of the informational prime beyond AA is consistent with AA being a prime that triggers a stereotype threat. However, it is also possible that the gender stereotype has no effect at all—that is, that AA encourages low-ability women by sufficiently increasing the marginal benefit of their effort while at the same time making the marginal benefit of extra effort not worth it for the high-ability women. This could explain the observed pattern in the effect of AA: positive for low-ability women and negative for high-ability women. Another possibility is that AA leads women to focus on single-sex competition, which may encourage low-ability women to compete, consistent with the finding that women compete more when they are only competing against other women (e.g., Gneezy et al. 2003), while at the same time reducing high-ability women's concern about the competition. We examine these alternative explanations next.

3.2.1 Optimal Effort Response

Although AA (weakly) increases women's absolute chance of winning the bonus, women's optimal effort is determined by the marginal effect of effort on their probability of winning.

[8] The information prime is about women's inferior performance relative to men on GRE, which is, on average, 15% lower.

Hence, to determine whether it is optimal for women to increase their effort in response to AA, we need to examine how the effect of effort on their probability of winning the bonus changes with AA. That is, if greater effort means a greater increase in probability of winning under AA, we would expect higher effort (and therefore scores) under AA; if greater effort means a lower increase in the probability of winning under AA, we would expect lower effort under AA. The formal condition (see the appendix) shows that higher effort under AA is optimal when the marginal change in the probability of winning without AA ($p'(e; AA = 0)$) is lower than the marginal change in the probability of winning against the other woman, weighted by the chance that the single-sex competition ends up determining who wins the bonus. Since the single-sex competition is likely more important for women with lower ability, and since the marginal effect of effort on their probability of winning without AA is likely lower, the pattern of performance observed may simply be the optimal reaction to AA.

To ascertain whether this is the case, we have to determine whether the marginal change in probability due to greater effort is indeed higher for low-ability women and lower for high-ability women under AA. To do that, we calculate for each woman, on the basis of her first-round performance, whether she would win a competition against her group members. We repeat that exercise assuming that she solves one more question correctly (as a proxy for exerting more effort), and we take the difference. This is the marginal effect of effort on the probability of winning. We do this once under the assumption of no AA and once under the assumption of AA.

To test for equality of the marginal probability of winning, we use the bootstrap method and find that, overall, the marginal probability of winning is higher under AA and marginally significant with a p-value of 0.102. Once we split the analysis to examine high- and low-ability women separately, we find that for high-ability women the difference is zero, while for low-ability women the difference is positive and significant.[9] This calculation suggests that it is not optimal for women of any level of ability to reduce their effort.[10]

Although not optimal, it may still be that women exert less effort in response to AA and that this reduction in effort is more pronounced for high-ability women. To examine this hypothesis, we look at participants' self-reported effort during the study (ranging from 1 to 7, where 7 represents the highest effort), as indicated by their responses to our exit questionnaire. We run ordered probit regressions of those responses on whether the participant was assigned to the AA condition; the participant's first-round score; its interaction with ability; and the controls, including the number of questions attempted in the first round, the participant's age, and the participant's self-reported SAT scores (quantitative and verbal). Table 5A presents the results for women and Table 5B for men. Both tables report results from the two specifications considered before: one using the continuous measure of ability and the second using ability groups. We also note that only three individuals (two women and one man) reported a low effort level of 1 or 2. The tables, therefore, present the results both including these individuals (columns 1, 3, and 5) and excluding them (columns 2, 4, and 6). The results show that women exert more effort under AA, and this finding is robust to adding the interaction of the dummy variable of quota policy with ability (measured by first-round score) when the outliers are excluded. Including all women, including the two outliers, the effect is always positive albeit sometimes insignificant. For men, we find an insignificant effect of AA, whether or not the effect of AA is allowed to vary with ability and the outlier is excluded.

[9] Since we are using the round 1 score in our calculation, we determine high-/low-ability women by using their reported quantitative SAT scores. We use different thresholds for the high-/low-ability classifications: 770, 780, 790, and 800. The results are the same no matter which threshold is selected, and they are significant at the 5% significance level.

[10] Note that the marginal winning probability is calculated by assuming an additional question solved correctly, which is similar to the actual average difference in score between round 1 and round 2. That is, the marginal winning probabilities we calculated are relevant for the effort decision made in round 2.

Hence, even if AA changes women's perception of their probability of winning, it does not decrease their self-reported effort in response; in fact, the evidence suggests that they increase their effort across the board.[11]

To recap: it is not optimal for women to reduce their effort in response to AA, and indeed, the women in our study did not seem to do that. Thus, a reduction in effort does not appear to explain the lowered performance of high-ability women under AA.

3.2.2 Women-Only Competition

To look into the possibility that single-sex competition reduces high-ability women's concern and effort while encouraging low-ability women to compete, we run an additional condition in which the second-round competition is between two women. Other than the different group size and the single-sex composition, this condition is identical to the other conditions (same task and bonus for the winner). Thirty-four women from the same subject pool participated in this condition.

Using the data from this condition, we test whether women's performance in the paired, women-only condition differs from women's performance in the control group, where subjects are assigned to mixed-sex groups and are not subject to AA. Table 6 shows no significant difference between the two groups: in both conditions, neither the main effect nor its interaction with ability is significant. In other words, the effect of AA on women's performance does not seem to be due to a shift in focus from a mixed-sex group to a single-sex group.

Taken together, the findings are inconsistent with higher-ability women exerting less effort in response to AA, either because they are competing against another woman or because they have higher chances of winning. Hence, the negative effect of AA on high-ability women supports the hypothesis that AA policy evokes a stereotype-threat effect. It is interesting to note that the result—that the negative stereotype threat effect is evident only among high-ability women—is consistent with Preckel et al. (2008), who found evidence that the greatest gender

[11] Given that the effort measure is self-reported, it may be affected by the condition. Hence, the results suggest women that do not believe they should exert less effort under AA.

gap in confidence in one's ability is present among the most gifted children. This is also consistent with the stereotype threat literature, which suggests that the stereotype effect presents itself among those who highly identify with the domain of the task (e.g., Aronson et al. 1999, Steele et al. 2002, Wei 2009). In our study, high first-round ability is likely to be a good proxy for identification with the math domain. Hence, the stereotype threat theory predicts that the effect will concentrate among the high-ability women.

3.3. What Drives the Change in Performance?

Two factors together determine the score: the number of questions attempted and the success rate on those questions. To test whether AA affects both factors or only one, we analyze each factor separately.

3.3.1. Number of Questions Attempted

Examining the effect of AA on the number of questions that women attempt, we find an insignificant negative main effect of AA (Table 7, columns 1 and 2). For men, we find a positive but insignificant main effect of AA (Table 7, columns 3 and 4). For both men and women, we find that this effect is unrelated to the ability of the participant.

3.3.2. Success Rate

The analysis of the number of questions attempted does not fully explain the observed effect of AA on women, so we turn to examining the success rate. Table 8, columns 1 and 2, shows that the main effect of AA is positive and statistically significant (with a coefficient of 0.149, $p = 0.021$ in column 1; 0.120, $p = 0.025$ in column 2). We also find that this effect decreases with ability (with a coefficient of -0.013, $p = 0.097$ in column 1; -0.05, $p = 0.24$ in column 2). That is, the effect of AA on women is reflected in their success rate in answering the questions.

The pattern we find matches the pattern we observe in the analysis of scores. Namely, while AA does not significantly change the number of questions attempted, it positively affects the success rate of low-ability women and negatively affects the success rate of high-ability women. Together, the number of correctly answered questions (not shown) significantly increases on average by 1.406 questions (p = 0.033), and this positive effect declines with ability at the rate of −0.192 (p = 0.047); when using ability groups, the number of correctly solved questions increases by 1.216 (p = 0.014) and declines at the rate of −0.923 (p = 0.076).

For men (Table 8, columns 3 and 4), we find that the main effect of AA and its interaction with ability is small and insignificant.

3.4. Round 3

An interesting question is whether the effect of AA that we find in the second round is also present in round 3, where incentives are back to piece-rate. Using the bootstrap method to construct the diff-in-diff measure we used in section 3.1., we find that the performance change between round 3 and round 2 for high-ability women strongly mirrors that between round 2 and round 1 (see Figure 2), such that high-ability women perform similarly in round 3 and round 1. That is, we find no effect of AA on high-ability women's performance when AA is no longer implemented and under the same incentive structure (round 3 vs. round 1).

For mid-ability women, while the effect of AA on performance in round 2 is insignificant, the change between round 3 and round 2 is significantly negative (observed in 94% of the subsamples), leading to a negative effect of AA between round 3 and round 1. For low-ability women, in contrast, although their performance decreases between round 3 and round 2, this effect is insignificant and so the overall effect of AA between round 3 and round 1 seems to be a positive one.

Although the main focus of this paper is on the effect of AA on performance at the stage where the policy is implemented, the results of round 3 indicate possibly a more complex effect of such policies in the long run. This calls for further investigation.

4. Conclusion

We find that AA affects the performance of women on quantitative GRE questions in an incentivized and competitive environment. We observe a positive and significant effect on women with lower baseline ability but a negative and significant effect on women with higher baseline ability. Given the marginal return to effort across conditions, reducing effort is not optimal for women of any ability level, and self-reported effort among women in our sample confirms that AA does not lead to intentional effort reduction. We also find no evidence that single-sex competition between women explains the negative effect on the performance of high-ability women. Lastly, we find that AA in and of itself has a similar effect on performance whether or not it is accompanied by a direct stereotype prime. These results are therefore consistent with AA acting as a stereotype prime that leads to the unintended negative consequence of impairing performance of the protected group.

With the substantial interest around the world in AA policies aimed at advancing women, it is important to understand fully the potential effect of such policies. This paper provides first evidence that AA policies have stereotype threat effects. We hope that future work would examine this issue further to provide additional evidence on the occurrence and magnitude of the stereotype threat effect of AA polices.

References

Aronson, Joshua, Michael J. Lustina, Catherine Good, and Kelli Keough. 1999. When white men can't do math: necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology* 35: 29-46.

Balafoutas, Loukas, and Matthias Sutter. 2012. Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science* 335:579–582.

Calsamiglia, Caterina, Jörg Franke, and Pedro Rey-Biel. 2013. The incentive effects of affirmative action in a real-effort tournament. *Journal of Public Economics* 98:15–31.

Catalyst Census. 2013. Fortune 500 Women Board Directors. Available at: http://www.catalyst.org/system/files/2013_catalyst_census_fortune_500_women_board_director.pdf.

Croson Rachel, and Uri Gneezy. 2009. Gender Differences in Preferences. *Journal of Economic Literature* 47(2): 448-474.

European Commission – Directorate-General for Justice, 2012. Women in Economic Decision-making in the EU: Progress Report.
Available at: http://ec.europa.eu/justice/gender-equality/files/women-on-boards_en.pdf.

Fryer, Roland, Steven Levitt, John List. 2008. Exploring the impact of financial incentives on stereotype threat. *American Economic Review: Papers and Proceedings*.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini. 2003. Performance in competitive environments: Gender differences. *Quarterly Journal of Economics* 118:1049–1074.

Hyde, Janet Shibley, Elizabeth Fennema, and Susan J. Lamon. 1990. Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin* 107(2):139–155.

Meeting of the OECD Council at Ministerial level, Paris 25-26 May 2011, Report of the Gender Initiative: Gender Equality in Education, Employment and Entrepreneurship. Available at: http://www.oecd.org/education/48111145.pdf.

National Association for Law Placement press release, December 11, 2013. Representation of Women Associates Falls for Fourth Straight Yeas as Minority Associates Continue to Make Gains – Women and Minority Partners Continue to Make Small Gains.

Niederle, Muriel, and Lise Vesterlund. 2007. Do Women Shy Away from Competition? Do Men Compete Too Much? *The Quarterly Journal of Economics* 122(3): 1067-1101.

Niederle, Muriel, and Lise Vesterlund. 2010. Explaining the Gender Gap in Math Test Scores: The Role of Competition. *The Journal of Economic Perspectives* 24(2): 129-144.

Niederle, Muriel, Carmit Segal, Lise Vesterlund. 2013. How costly is diversity? Affirmative action in light of gender differences in competitiveness. *Management Science* 59(1): 1-16.

Preckel Franzis, Thomas Goetz, Reinhard Pekrun, and Michael Kleine. 2008. Gender Differences in Gifted and Average-Ability Students: Comparing Girls' and Boys' Achievement, Self-Concept, Interest, and Motivation in Mathematics. *Gifted Child Quarterly* 52(2): 146-59.

Schotter, Andrew, and Keith Weigelt. 1992. Asymmetric tournaments, equal opportunity laws, and affirmative action: Some experimental results. *Quarterly Journal of Economics* 107(2):511–539.

Shih, Margaret, Todd L. Pittinsky, and Nalini Ambady. 1999. Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science* 10:80–83.

Spencer, Steven J., Claude M. Steele, Diane M. Quinn. 1999. Stereotype threat and women's math performance. *Journal of Experimental Social Psychology* 35:4–28.

Steele, Claude M. and Joshua Aronson. 1995. Contending with a stereotype: African-American intellectual test performance and stereotype threat. *Journal of Personality and Social Psychology* 69:797–811.

Steele, Claude M., Steven J. Spencer, and Joshua Aronson. 2002. Contending With Group Image: The Psychology of Stereotype and Social Identity Threat. *Advances in Experimental Social Psychology* 34: 379-440.

Stricker, Lawrence J. 1998. Inquiring about examinees' ethnicity and sex: Effects on AP Calculus AB examination performance. Collage Board Report No. 98-1. ETS Report No. 98-5.

Sutter, Matthias and Daniela Rützler. 2010. Gender Differences in Competition Emerge Early in Life. IZA Discussion Paper No. 5015.

U.S. Department of Commerce, Economics and Statistics Administration. 2011. Women in STEM: A Gender Gap to Innovation.
Available at: http://www.esa.doc.gov/Reports/women-stem-gender-gap-innovation.

Wei, Thomas. 2009. Under what conditions? Stereotype threat and prime attributes. Working Paper. Available at http://www.people.fas.harvard.edu/~twei/papers/sthreat_exper.pdf.

Wei, Thomas. 2012. Sticks, stones, words, and broken bones: New field and lab evidence on stereotype threat. *Educational Evaluation & Policy Analysis* 34:465.

Table 1: Summary Statistics

		Male	Female	Diff.
		(mean)	(mean)	
Round 1	Score	6.36	5.65	$p<0.10$ one sided
	# Questions	8.91	8.32	$p<0.10$ one sided
	Ratio correct	0.76	0.72	insignificant
Round 2	Score	7.31	6.45	$p<0.10$ two-sided
	# Questions	11.50	9.89	$p<0.01$ two-sided
	Ratio correct	0.69	0.71	insignificant
Round 3	Score	7.54	7.06	insignificant
	# Questions	12.20	11.26	$p<0.05$ two-sided
	Ratio correct	0.68	0.68	insignificant

* $p<0.1$, ** $p<0.05$, *** $p<0.01$

Table 2: The Effect of Affirmative Action on Performance

	(1) Female	(2) Female	(3) Male	(4) Male
Affirmative Action (AA)	0.436	1.889**	0.566	1.051
	(0.432)	(0.795)	(0.532)	(0.904)
Score in 1st round (Score R1)	-0.723***	-0.572***	-0.372***	-0.340**
	(0.078)	(0.093)	(0.141)	(0.137)
(AA)x(Score R1)		-0.230**		-0.067
		(0.108)		(0.128)
No. Questions in 1st Rnd	0.442***	0.471***	0.187	0.171
	(0.069)	(0.075)	(0.132)	(0.130)
Age	0.000	-0.052	-0.173	-0.160
	(0.106)	(0.111)	(0.180)	(0.177)
SAT Quantitative	0.018***	0.017***	0.022***	0.023***
	(0.003)	(0.003)	(0.006)	(0.006)
SAT Verbal	0.002	0.002	0.002	0.002
	(0.003)	(0.003)	(0.004)	(0.004)
Constant	-13.242***	-13.153***	-13.356**	-14.058**
	(3.974)	(3.931)	(5.905)	(6.019)
N	123	123	118	118
R^2	0.39	0.42	0.22	0.22

Notes: WLS regressions. Dependent variable: difference in score between round 2 and round 1. Robust standard errors are reported in parentheses. * indicates significance at the 10-percent level; ** indicates significance at the 5-percent level; *** indicates significance at the 1-percent level.

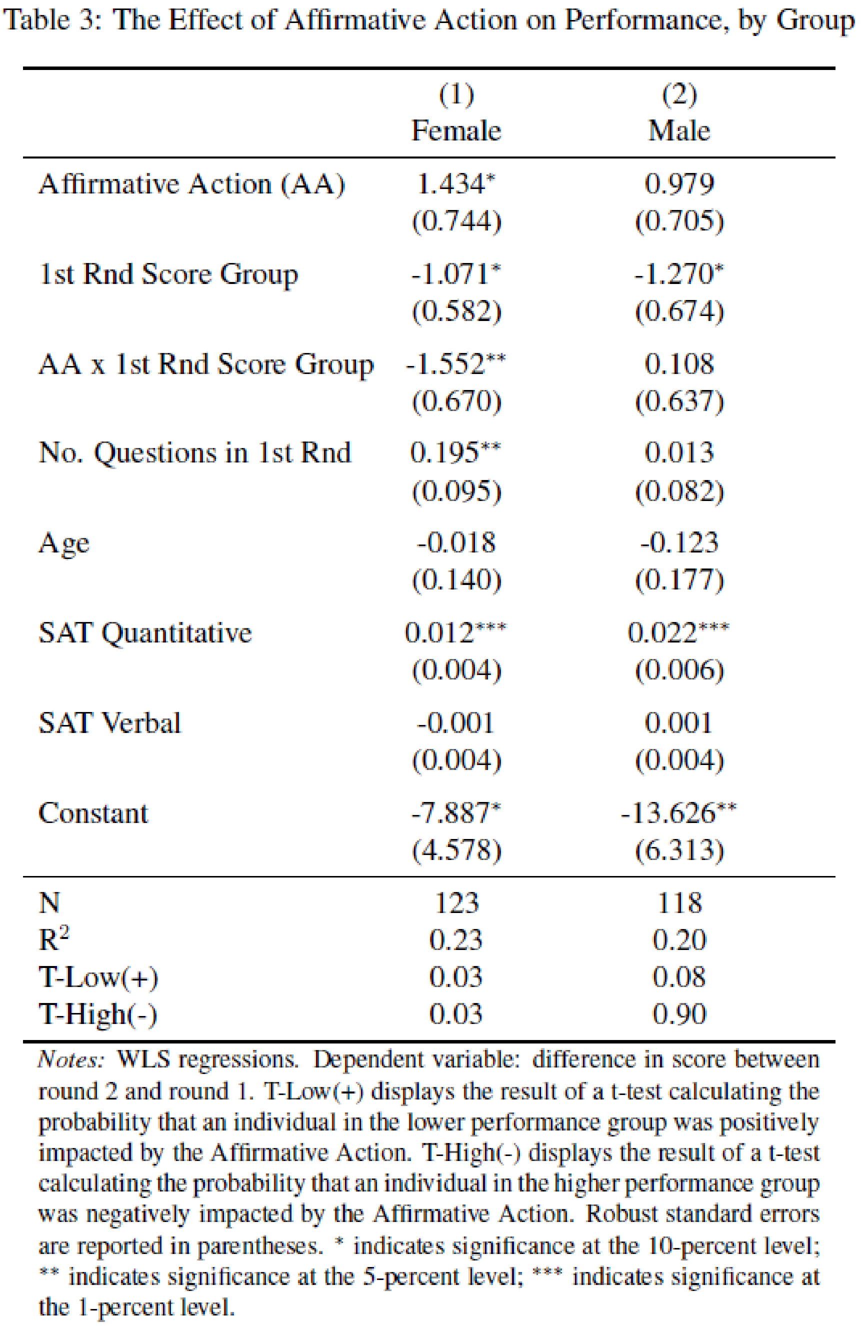


Table 3: The Effect of Affirmative Action on Performance, by Group

	(1) Female	(2) Male
Affirmative Action (AA)	1.434*	0.979
	(0.744)	(0.705)
1st Rnd Score Group	-1.071*	-1.270*
	(0.582)	(0.674)
AA x 1st Rnd Score Group	-1.552**	0.108
	(0.670)	(0.637)
No. Questions in 1st Rnd	0.195**	0.013
	(0.095)	(0.082)
Age	-0.018	-0.123
	(0.140)	(0.177)
SAT Quantitative	0.012***	0.022***
	(0.004)	(0.006)
SAT Verbal	-0.001	0.001
	(0.004)	(0.004)
Constant	-7.887*	-13.626**
	(4.578)	(6.313)
N	123	118
R^2	0.23	0.20
T-Low(+)	0.03	0.08
T-High(-)	0.03	0.90

Notes: WLS regressions. Dependent variable: difference in score between round 2 and round 1. T-Low(+) displays the result of a t-test calculating the probability that an individual in the lower performance group was positively impacted by the Affirmative Action. T-High(-) displays the result of a t-test calculating the probability that an individual in the higher performance group was negatively impacted by the Affirmative Action. Robust standard errors are reported in parentheses. * indicates significance at the 10-percent level; ** indicates significance at the 5-percent level; *** indicates significance at the 1-percent level.

Table 4: Affirmative Action and Informational Prime

	(1)	(2)
	Female	Male
Affirmative Action (AA)	1.696**	1.350
	(0.755)	(0.830)
Info	-0.519	-0.616
	(0.701)	(0.958)
1st Rnd Score Group	-1.071*	-1.277*
	(0.586)	(0.683)
AA x 1st Rnd Score Group	-1.601**	-0.048
	(0.741)	(0.784)
Info x 1st Rnd Score Group	0.064	0.257
	(0.753)	(0.842)
No. Questions in 1st Rnd	0.196**	0.021
	(0.096)	(0.082)
Age	-0.026	-0.135
	(0.135)	(0.181)
SAT Quantitative	0.012***	0.021***
	(0.004)	(0.006)
SAT Verbal	-0.001	0.001
	(0.004)	(0.004)
Constant	-7.644*	-12.898*
	(4.526)	(6.811)
N	123	118
R^2	0.23	0.21

Notes: WLS regressions. Dependent variable: difference in score between round 2 and round 1. Robust standard errors are reported in parentheses. * indicates significance at the 10-percent level; ** indicates significance at the 5-percent level; *** indicates significance at the 1-percent level.

Table 5A: The Effect of Affirmative Action on Effort, Females

	(1)	(2)	(3)	(4)	(5)	(6)
Affirmative Action (AA)	0.405*	0.473**	0.549	0.878**	0.497	0.668*
	(0.211)	(0.221)	(0.374)	(0.404)	(0.318)	(0.353)
Score in 1st round (Score R1)	0.036	0.041	0.052	0.086		
	(0.049)	(0.049)	(0.052)	(0.057)		
(AA)x(Score R1)			-0.026	-0.071		
			(0.051)	(0.055)		
1st Rnd Score Group					0.187	0.392
					(0.247)	(0.270)
AA x 1st Rnd Score Group					-0.093	-0.200
					(0.256)	(0.279)
No. Questions in 1st Rnd	-0.017	-0.005	-0.014	0.006	-0.007	-0.002
	(0.048)	(0.046)	(0.049)	(0.045)	(0.037)	(0.037)
Age	0.017	0.009	0.012	-0.006	0.012	-0.001
	(0.064)	(0.068)	(0.067)	(0.073)	(0.066)	(0.071)
SAT Quantitative	0.005**	0.002	0.005**	0.002	0.005**	0.002
	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.002)
SAT Verbal	-0.002	-0.003	-0.002	-0.003	-0.002	-0.003
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
N	123	121	123	121	123	121

Notes: Ordered probit regressions. Dependent variable: self-reported effort level. Robust standard errors are reported in parentheses. * indicates significance at the 10-percent level; ** indicates significance at the 5-percent level; *** indicates significance at the 1-percent level.

Table 5B: The Effect of Affirmative Action on Effort, Males

	(1)	(2)	(3)	(4)	(5)	(6)
Affirmative Action (AA)	0.083	0.142	0.205	0.293	0.082	0.152
	(0.198)	(0.202)	(0.420)	(0.455)	(0.272)	(0.283)
Score in 1st round (Score R1)	0.129**	0.147***	0.139**	0.159**		
	(0.054)	(0.056)	(0.064)	(0.068)		
(AA)x(Score R1)			-0.019	-0.023		
			(0.063)	(0.067)		
1st Rnd Score Group					0.349*	0.368*
					(0.206)	(0.212)
AA x 1st Rnd Score Group					-0.160	-0.204
					(0.233)	(0.239)
No. Questions in 1st Rnd	-0.094*	-0.112**	-0.097**	-0.116**	-0.030	-0.034
	(0.050)	(0.053)	(0.049)	(0.052)	(0.043)	(0.045)
Age	0.005	0.000	0.008	0.003	-0.011	-0.017
	(0.063)	(0.065)	(0.063)	(0.065)	(0.065)	(0.066)
SAT Quantitative	0.003	0.002	0.003	0.002	0.004*	0.004
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
SAT Verbal	0.000	0.001	0.000	0.001	0.000	0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
N	118	117	118	117	118	117

Notes: Ordered probit regressions. Dependent variable: self-reported effort level. Robust standard errors are reported in parentheses. * indicates significance at the 10-percent level; ** indicates significance at the 5-percent level; *** indicates significance at the 1-percent level.

Table 6: Women-Only Competition

	(1)	(2)
	1st Rnd Score	1st Rnd Score Group
Women only	0.235	0.366
	(0.802)	(0.743)
Score in 1st round (Score R1)	-0.451***	
	(0.124)	
1st Rnd Score Group		-0.667
		(0.586)
(Women only)x(Score R1)	0.025	
	(0.095)	
(Women only)*(1st Rnd Score Group)		-0.248
		(0.592)
No. Questions in 1st Rnd	0.311**	0.074
	(0.122)	(0.078)
Age	-0.066	-0.082
	(0.120)	(0.146)
SAT Quantitative	0.012*	0.004
	(0.006)	(0.007)
SAT Verbal	0.003	0.001
	(0.003)	(0.004)
N	72	72
R^2	0.19	0.06

Notes: WLS regressions. Dependent variable: difference in score between round 2 and round 1. Robust standard errors are reported in parentheses. * indicates significance at the 10-percent level; ** indicates significance at the 5-percent level; *** indicates significance at the 1-percent level.

Table 7: Number of Questions

	(1) Female	(2) Female	(3) Male	(4) Male
Affirmative Action (AA)	-0.529	-0.085	1.211	0.631
	(0.904)	(0.673)	(0.911)	(0.778)
Score in 1st round (Score R1)	0.178		0.085	
	(0.149)		(0.143)	
(AA)x(Score R1)	-0.037		-0.064	
	(0.138)		(0.110)	
1st Rnd Score Group		1.252*		-0.364
		(0.654)		(0.662)
AA x 1st Rnd Score Group		-0.570		0.015
		(0.711)		(0.636)
No. Questions in 1st Rnd	0.526***	0.545***	0.632***	0.723***
	(0.197)	(0.148)	(0.150)	(0.085)
Age	0.091	0.074	-0.083	-0.088
	(0.249)	(0.244)	(0.150)	(0.143)
SAT Quantitative	0.009*	0.008**	0.010**	0.013**
	(0.005)	(0.004)	(0.005)	(0.005)
SAT Verbal	-0.005	-0.006	0.002	0.002
	(0.004)	(0.004)	(0.004)	(0.004)
N	123	123	118	118
R^2	0.60	0.61	0.67	0.68

Notes: WLS regressions. Dependent variables: number of questions answered in round 2. Robust standard errors are reported in parentheses. * indicates significance at the 10-percent level; ** indicates significance at the 5-percent level; *** indicates significance at the 1-percent level.

Table 8: Success Rate

	(1) Female	(2) Female	(3) Male	(4) Male
Affirmative Action (AA)	0.149**	0.120**	0.019	0.024
	(0.064)	(0.053)	(0.063)	(0.051)
Score in 1st round (Score R1)	0.019**		0.034***	
	(0.009)		(0.008)	
(AA)x(Score R1)	-0.013*		0.000	
	(0.008)		(0.007)	
1st Rnd Score Group		0.064		0.115***
		(0.040)		(0.043)
AA x 1st Rnd Score Group		-0.051		-0.045
		(0.043)		(0.040)
No. Questions in 1st Rnd	-0.001	0.002	-0.019**	-0.004
	(0.009)	(0.006)	(0.007)	(0.006)
Age	0.004	0.004	-0.013	-0.018
	(0.010)	(0.010)	(0.010)	(0.011)
SAT Quantitative	0.001***	0.002***	0.001***	0.002***
	(0.000)	(0.000)	(0.000)	(0.000)
SAT Verbal	0.001**	0.001**	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
N	123	123	118	118
R^2	0.42	0.41	0.47	0.42

Notes: OLS regressions. Dependent variables: success rate in round 2. Robust standard errors are reported in parentheses. * indicates significance at the 10-percent level; ** indicates significance at the 5-percent level; *** indicates significance at the 1-percent level.

Table 9: Patterns of the Affirmative Action Effect on Women

	(1) Round 2 - Round 1	(2) Round 3 - Round 2	(3) Round 3 - Round 1
Affirmative Action (AA)	1.434^{*}	-1.131^{t}	0.303
	(0.744)	(0.715)	(0.788)
1st Rnd Score Group	-1.071^{*}	1.550^{**}	0.479
	(0.582)	(0.603)	(0.523)
AA x 1st Rnd Score Group	-1.552^{**}	1.149^{*}	-0.403
	(0.670)	(0.663)	(0.602)
No. Questions in 1st Rnd	0.195^{**}	-0.206^{***}	-0.011
	(0.095)	(0.064)	(0.075)
Age	-0.018	-0.291^{t}	-0.309^{*}
	(0.140)	(0.190)	(0.171)
SAT Quantitative	0.012^{***}	-0.008^{*}	0.004
	(0.004)	(0.005)	(0.005)
SAT Verbal	-0.001	-0.000	-0.001
	(0.004)	(0.004)	(0.006)
Constant	-7.887^{*}	13.687^{**}	5.800
	(4.578)	(5.929)	(4.810)
N	123	123	123
R^2	0.23	0.26	0.06
T-Low(-)		0.06	0.65
T-Low(+)	0.03		0.35
T-High(-)	0.03		0.26
T-High(+)		0.11	0.74

Notes: WLS regressions. Dependent variables: difference in score between round 2 and round 1 (column 1), difference in score between round 3 and round 2 (column 2), difference in score between round 3 and round 1 (column 3). T-Low(+) displays the result of a t-test calculating the probability that an individual in the lower performance group was positively impacted by the Affirmative Action. T-Low(-) displays the result of a t-test calculating the probability that an individual in the lower performance group was negatively impacted by the Affirmative Action. T-High(-) displays the result of a t-test calculating the probability that an individual in the higher performance group was negatively impacted by the Affirmative Action. T-High(+) displays the result of a t-test calculating the probability that an individual in the higher performance group was positively impacted by the Affirmative Action. Robust standard errors are reported in parentheses. * indicates significance at the 10-percent level; ** indicates significance at the 5-percent level; *** indicates significance at the 1-percent level.

Figure 1

Females' [Round 2- Round 1] Score Difference, by AA

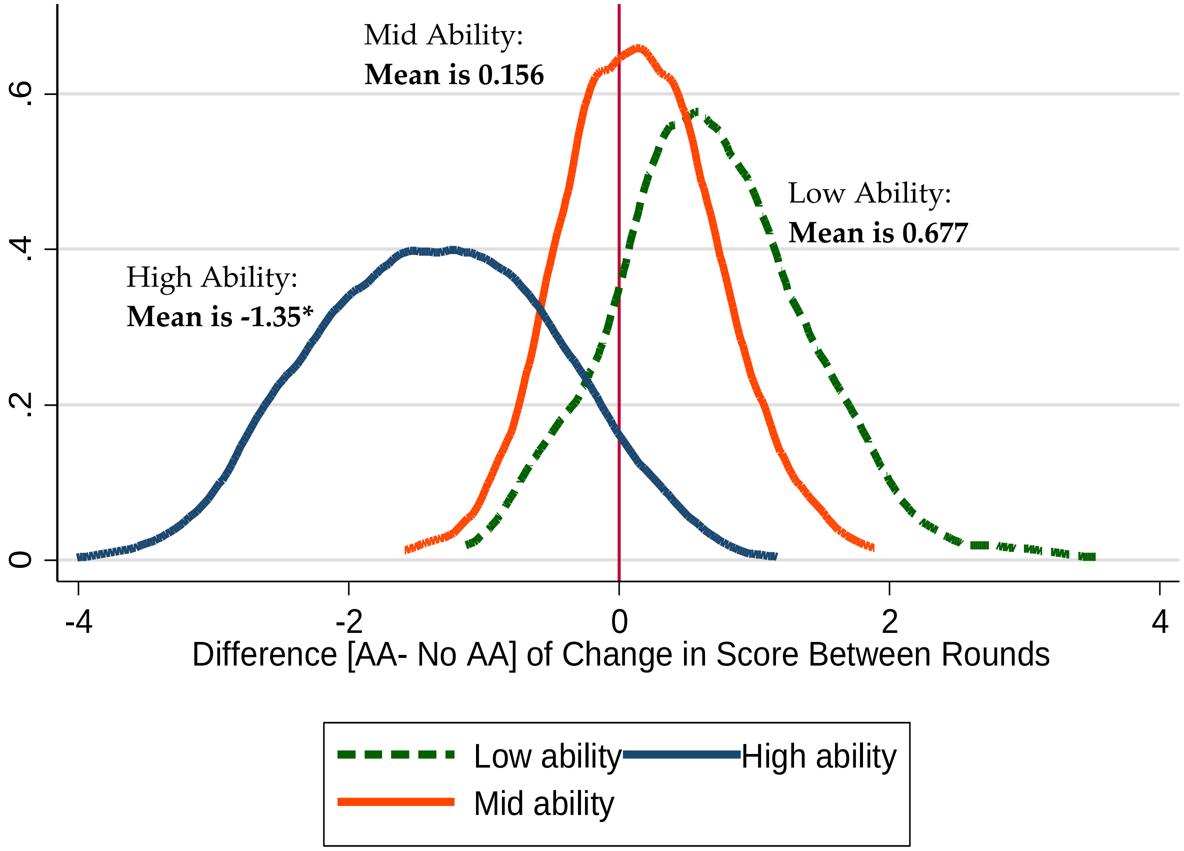
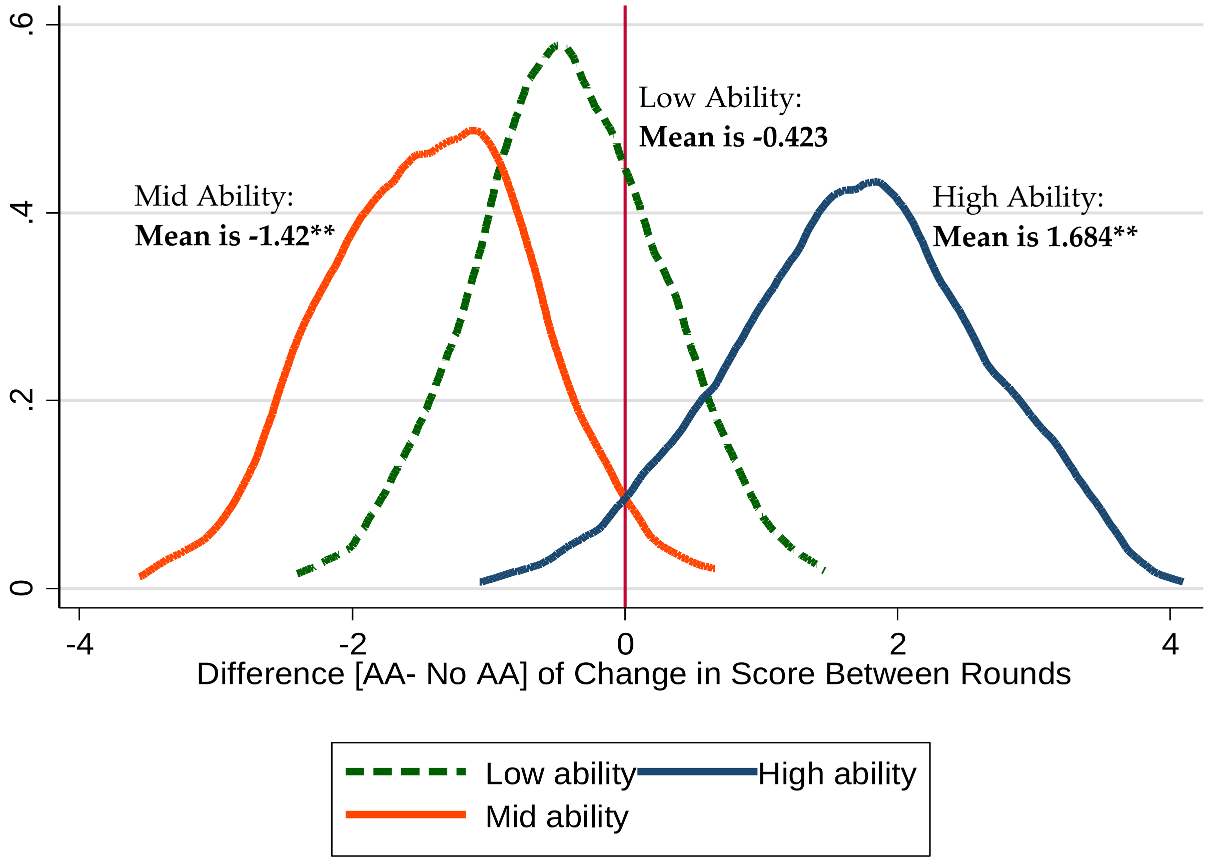


Figure 2

Females' [Round 3- Round 2] Score Difference by AA



Appendix:

If the model is

$$max_e\, p(e;AA)[Bonus + Reward(e) - c(e)] + [1 - p(e;AA)][Reward(e) - c(e)] = \\ max_e pe;AABonus +\ Rewarde - ce,$$

then the FOC is

$$p'(e;AA)Bonus +\ Reward'(e) = c'(e).$$

So, if $p'(e;AA=1) \geq p'(e;AA=0)$, we expect greater effort (weakly).

But if $p'(e;AA=1) < p'(e;AA=0)$, effort should decline.

Let define the following terms:

$p_1(e)$ – to be the probability of being the highest performer in the group of four.
$p_2(e)$ – to be the probability of being the second-highest performer in the group of four.
q – to be the probability that the two top performers in the group of four are men.
$w_1(e)$ – to be the probability of being the highest performer in the group of two women.

The probability of winning the bonus when AA=0 is

$$p(e;AA=0) = p_1(e) + [1 - p_1(e)]p_2(e).$$

The probability of winning the bonus when AA=1 is

$$p(e;AA=1) = p_1(e) + [1 - p_1(e)]p_2(e) + [1 - p_1(e)][1 - p_2(e)]qw_1(e).$$

The change in probability as effort (e) increases when AA is equal to 0:

$$\begin{aligned} p'(e;AA=0) &= p_1'(e) + [1 - p_1(e)]p_2'(e) - p_1'(e)p_2(e) \\ &= p_1'(e)[1 - p_2(e)] + [1 - p_1(e)]p_2'(e) \end{aligned}$$

The change in probability as effort (e) increases when AA is equal to 1:

$$\begin{aligned} p'(e;AA=1) = {} & p_1'(e)[1 - p_2(e)] + [1 - p_1(e)]p_2'(e) - p_1'(e)[1 - p_2(e)]qw_1(e) - \\ & -[1 - p_1(e)]p_2'(e)qw_1(e) + [1 - p_1(e)][1 - p_2(e)]qw_1'(e) = \\ & = p_1'(e)[1 - p_2(e)][1 - qw_1(e)] + [1 - p_1(e)]p_2'(e)[1 - qw_1(e)] + \\ & +[1 - p_1(e)][1 - p_2(e)]qw_1'(e). \end{aligned}$$

Which one is greater? $p'(e;AA=0) \lesseqgtr p'(e;AA=1)$?

$$\begin{gathered} p'(e;AA=0) \lesseqgtr p'(e;AA=1) \Leftrightarrow \\ p_1'(e)[1-p_2(e)]+[1-p_1(e)]p_2'(e) \lesseqgtr \\ p_1'(e)[1-p_2(e)][1-qw_1(e)]+[1-p_1(e)]p_2'(e)[1-qw_1(e)] \\ +[1-p_1(e)][1-p_2(e)]qw_1'(e). \end{gathered}$$

That is,

$$p_1'(e)[1-p_2(e)]w_1(e)+[1-p_1(e)]p_2'(e)w_1(e) \lesseqgtr [1-p_1(e)][1-p_2(e)]w_1'(e).$$

This can be seen as

$$(1) \quad \underbrace{p_1'(e)[1-p_2(e)]+[1-p_1(e)]p_2'(e)}_{p'(e;AA=0)} \lesseqgtr \underbrace{[1-p_1(e)][1-p_2(e)]\frac{w_1'(e)}{w_1(e)}}_{\substack{\text{Percentage increase in the chance to win} \\ \text{against the other woman multiplied by the} \\ \text{probability of not being in the top two.}}}$$

Or:

$$(2) \quad \frac{p_1'(e)}{[1-p_1(e)]}+\frac{p_2'(e)}{[1-p_2(e)]} \lesseqgtr \frac{w_1'(e)}{w_1(e)}\,.$$

For a given effort level and for high-ability women, their chance of not being among the top two is lower than it is for low-ability women. Also, the high-ability women have higher chance of winning against the other woman. So the weight on the RHS in (1) is lower, and it is likely that the LHS>RHS. For the low-ability women, the opposite holds.