

NBER WORKING PAPER SERIES

THE HETEROGENEOUS EFFECT OF AFFIRMATIVE ACTION ON PERFORMANCE

Anat Bracha
Alma Cohen
Lynn Conell-Price

Working Paper 25322
<http://www.nber.org/papers/w25322>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2018

We would like to thank Uri Gneezy, Muriel Niederle, Tali Regev, Analia Schlosser, Lise Vesterlund 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, the Federal Reserve System, or the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2018 by Anat Bracha, Alma Cohen, and Lynn Conell-Price. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

The Heterogeneous Effect of Affirmative Action on Performance
Anat Bracha, Alma Cohen, and Lynn Conell-Price
NBER Working Paper No. 25322
December 2018
JEL No. C91,I28,J16,J78,K19,K31

ABSTRACT

This paper experimentally investigates the effect of gender-based affirmative action (AA) on performance in the lab, focusing on a tournament environment. The tournament is based on GRE math questions commonly used in graduate school admission, and at which women are known to perform worse on average than men. We find heterogeneous effect of AA on female participants: AA lowers the performance of high-ability women and increases the performance of low-ability women. Our results are consistent with two possible mechanisms—one is that AA changes incentives differentially for low- and high-ability women, and the second is that AA triggers stereotype threat.

Anat Bracha
Research Department
600 Atlantic Ave
Boston, MA 02210
anat.bracha@bos.frb.org

Lynn Conell-Price
5000 Forbes Avenue
BP 208
Pittsburgh, PA 15213
USA
lconellp@andrew.cmu.edu

Alma Cohen
Harvard Law School
1525 Massachusetts Avenue
Cambridge, MA 02138
and NBER
alcohen@law.harvard.edu

1. Introduction

Affirmative action (AA) is a hotly debated policy, with passionate arguments in favor of and against it. While the arguments center around issues of fairness, discrimination, and the interplay of AA and stereotypes (e.g., APA 2012, Schuck 2002), there is also the question of the extent to which AA improves, or worsens, outcomes for its intended beneficiaries (Sander 2010, Imbens et al. 2012). On this point, there is an underexplored empirical question: how does AA affect the performance of its intended beneficiaries? College admission processes using AA could, for instance, affect performance on the SAT exams or investments in education at earlier stages of life.

This paper seeks to shed light on this question by examining the effect of gender-based AA on women's performance on math admission exams in a lab experiment. We add to the literature by examining the effect of AA on performance in tests like those used to make high-stakes admissions decisions and are the first to find experimental evidence of its heterogeneous effect. Specifically, the results of the lab experiment find a positive effect of AA on the performance of low-ability women but a negative effect of AA on the performance of high-ability women, suggesting a new nuance in the AA debate. That is, not only should researchers and policymakers ask whether AA benefits disadvantaged groups, they should also consider the possibility that AA may differentially affect high-performance and low-performance individuals in those groups.

We chose to focus on gender-based AA policies as this is a timely topic that attracts substantial interest and policy debate. It is well known that although women today represent half of all college-educated individuals, they are still substantially underrepresented in many selective, high-level professional positions and occupations, such as corporate directors and top executives¹, partners in US law firms, and holders of science, technology, engineering, and math (STEM) jobs in the Organization for Economic Co-operation and Development (OECD) countries (see Catalyst Census: Fortune 500 Women Board Directors 2013 and European Commission 2012,

¹ Women constitute less than 20% of corporate directors, and their fraction of top executive positions in large public companies is below 5%.

National Association for Law Placement 2013, Meeting of the OECD Council at Ministerial Level 2011 and, for the United States, the U.S. Department of Commerce, Economics and Statistics Administration 2011).

This underrepresentation of women in certain professional groups and positions has led to a strong interest in AA policies. For example, in 2012 the European Commission proposed legislation to ensure that, by 2020, 40% of nonexecutive directors will be women (European Commission 2012). Adding to this are recent studies in economics that argue that underrepresentation of women in certain professions may be due, in part, to women's avoidance of competition, especially in mixed-sex settings, and AA could therefore help remedy the impacts of this aversion to competition (see, e.g., Gneezy et al. 2003, Niederle and Vesterlund 2007, Niederle and Vesterlund 2010, Sutter and Rützler 2010, Niederle et al. 2013, Balafoutas and Sutter 2012, and a review by Croson and Gneezy 2009).

We chose to examine performance on the GRE math questions since it is used in actual admission decisions, and because it is a challenging task on which women are known (and documented) to underperform compared with men (Willingham & Cole, 1997). Further, participants in our sample have a good understanding of their relative performance on this type of exam, since they are students who took the SAT, which contains very similar math questions.

The first two properties of the GRE math questions (challenging problems, and known underperformance of women) are essential to understanding the overall effect of AA. This is because while AA may curb the effects of women's noncompetitive tendencies, it may also affect their incentives to exert effort (positively or negatively) and influence their mindsets. In the context of admission exams, AA could increase the return to effort for lower-performing women by making admission a more achievable goal while at the same time decreasing the return to effort for higher-performing women who already expected admission. However, this effect is only likely when participants have a good understanding of their relative performance. For instance, a high-ability woman who may choose to reduce her efforts under AA when she knows she is of high relative ability would not necessarily choose to reduce her effort if she were not

aware of her relative ability. As for mindset, the theory of stereotype threat predicts that reminding people of a negative stereotype about their group can adversely affect their objective performance (Steele 1997, Steele & Aronson 1995, Steele et al., 2002; in economics see for example Hoff and Pandey 2006). While AA may trigger stereotype threat by making salient a negative stereotype about a group's performance in the domain that the policy targets (Leslie et al. 2014), this mindset effect, if it exists, appears only in performing challenging tasks (Spencer et al. 1999). Since in many real-life examples these properties hold, it is important to examine the effect of AA in such settings.

We use a lab experiment to examine the effect of AA on performance. Participants were assigned to groups of four, two men and two women each, and were asked to solve math GRE questions. Performance was incentivized using piece-rate payments and a bonus. In the control group, the bonus was awarded to the top two performers regardless of gender, while in the AA condition it was required that at least one woman receive the bonus. We then compared the individuals' performance (i.e., their piece-rate payment) across the different conditions and found that women with low baseline ability perform significantly better in the presence of an AA policy, whereas women with high baseline ability perform significantly worse.² These results hold whether we use linear or nonlinear specifications and whether we use a standard regression analysis or the bootstrapping method.

Our experimental study differs from past experimental studies of AA in the properties of the task that we use. In the context of gender-based AA, for example, Niederle et al (2013) study the effect of gender-based AA on the willingness to compete using the task of summing up five 2-digit numbers. This task is simple, with no known gender difference, participants are unlikely to know their relative performance due to the novelty of the task, and performance on it does not respond to incentives. These properties of the task are suitable for isolating competition preferences,

² Once participants were assigned to groups, those in the AA conditions were informed of the AA policy before they began solving the GRE questions. That is, the performance we analyze is their GRE performance while knowing that AA will be used at the end of the round to determine who gets the bonus, much like a student who takes the SAT exam and expects AA to affect the college admission process.

however, it cannot uncover potential effects on performance, such as incentive and mindset effects. The closest experimental work to our own, by Calsamiglia et al. (2013), also focuses on the effect of AA on performance. They find no negative effect of AA on the performance of either group (the advantaged or disadvantaged subjects). However, in their study the disadvantaged subjects have no prior knowledge of either their absolute or relative ability on the task used, and there is no ingrained stereotype associated with the task or the disadvantaged group, in contrast with real-world applications of AA that we attempt to capture.³

We identify three potential mechanisms that could explain our results. One is the incentive effect of AA, i.e., the increased chance of women winning the competition under AA, holding effort levels constant. While this effect increases the marginal return to effort (and hence optimal effort exertion) for lower ability women, it may lower the marginal return to effort for the highest ability women. Intuitively, this is because the competition becomes less fierce under AA and, as a result, the highest-ability women may not need to exert as much effort as without AA to secure their win. Importantly, even if the incentive effect is positive for all women, it will be lower for higher ability women than their low ability counterparts. Second, stereotype threat may reduce women's performance by triggering a negative mindset. Notably, if stereotype threat is induced for all women, its combination with changes in incentives will still lead to a more negative net effect (or smaller positive net effect) for higher ability than for lower ability women because of the difference in the offsetting incentive effect. Finally, single-sex as opposed to mixed-sex competition may also affect effort provision. This is relevant because gender-based AA may change women's perception of the competition from a mixed-sex to a single-sex competition, in which case we would expect AA to help boost women's performance (e.g., Gneezy et al. 2003). While we reject the single-sex competition mechanism using an experimental treatment, we cannot disentangle the incentive effect from the stereotype threat effect through our design. Our

³ Calsamiglia et al. (2013) examine school-based rather than gender-based AA, with children from two primary schools in Spain competing on Sudoku puzzles. The children across the two schools were similar except that one of the two schools had the advantage of prior Sudoku training. In light of this difference, the disadvantaged children, those who did not get Sudoku training prior to the experiment, were subject to AA.

findings suggest that if both mechanisms are at work, the incentive effect among low-ability women must be positive and larger than the stereotype threat effect. The negative effect of AA on the performance of high-ability women may be due to a positive incentive effect that is smaller than the negative stereotype threat effect, or to an incentive effect alone that is already negative. Disentangling the two effects is left for future work.

The findings indicate that AA may have unintended negative effects on test performance that are relevant to the assessment of AA policies, and that these effects may vary for people of different underlying ability. This makes the overall welfare impact unclear, demanding further work on this subject.

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 the effect of gender-based AA on performance, we employed a between-subject experimental design in which 248 participants were randomly assigned to gender-based AA in a competitive setting with incentivized performance. Participants were asked to solve math GRE questions.⁴ We chose such questions because (1) men's average performance on math GREs is known to be better than women's; (2) this setup resembles real-life situations where AA is introduced to counteract existing disadvantages; (3) participants have a good idea of their relative ability based on experience from previously taking the SAT; and (4) AA applied to a disadvantage group, when the task is challenging and important to the subjects, has the potential to trigger a stereotype threat effect.

⁴ The exam uses multiple-choice questions with five possible answers from which the examinee must select one. The questions were selected from previous versions of the actual GRE which are published after their use. All participants saw the same questions in the same order.

The experiment used three 10-minute rounds, during which participants were asked to answer as many questions as they could (there were enough questions in each round such that none of the participants ran out of questions). To measure performance and, in turn, examine the effect of AA, we calculated participants' performance score in each round. The score increased by one point with every correctly answered question and decreased by a quarter of a point with every incorrectly answered question.⁵ There was no feedback on performance given to participants until the end of the experiment.

Effort was incentivized in each round. In the first round, we used noncompetitive piece-rate incentives, with each point of the score earning a dollar. That is, each correctly answered question yielded a dollar and each incorrectly answered question reduced earnings by a quarter. At this point, participants were not aware of any AA policy and we can therefore use the round 1 score as a proxy for ability.

In the second round, which is the main focus of our analysis, participants were randomly assigned to a group of two men and two women.⁶ In this round, the payment consisted of two types of incentives: (1) a noncompetitive piece-rate incentive, as in round 1, and (2) a competitive incentive, with participants competing for an additional \$10 bonus. This reward structure is somewhat analogous to admission, as the piece rate can be thought of representing the GRE score and the additional bonus as being admitted or not. The rules for winning the bonus depended on the experimental condition to which each group was assigned. Specifically, groups were randomly assigned to one of two conditions: a control condition (No AA) and an AA condition (AA). Under the No AA condition, the two highest-performing group members won the tournament, and each received a \$10 bonus. Under the AA condition, the two highest-scoring

⁵ The current scoring rule for the GRE does not directly penalize incorrect responses in this manner. However, incorrect responses still indirectly hurt one's score by affecting the difficulty (and point value) of subsequent questions that the "Computerized Adaptive Test" (CAT) system presents. A way to capture this relationship in a lab setting is with a direct penalty, which is also consistent with past scoring rules of the GRE, as described in GRE test preparation books—see for example the 2009 Princeton Review "Cracking the GRE" and Henry George Stratakis-Allen's 2007 book "The complete Idiot's guide to Acing the GRE".

⁶ While all group members were present in the lab at the same time, it was not possible for them to identify who of the other people in the lab were in their group. Participants were not allowed to communicate with each other, and since they were not aware of group membership, communication was also irrelevant.

group members received a \$10 bonus subject to a gender quota that required at least one woman to receive the bonus. Thus, if the two participants with the highest scores were men, the man with the highest (overall) score and the woman with the highest (female) score would earn the bonus. However, if two women earned the highest scores they would both receive the bonus. The AA policy was explained to participants assigned to the AA condition at the outset of round 2, before they began solving GRE questions in this round. Two versions of the AA condition were used: AA without information (AA-no-I) and AA with information (AA-I). The two versions were identical except that, just before beginning the task, participants in the AA-I version received information on the gender gap in math GRE performance.

These two versions of the AA condition were designed a priori to disentangle two possible mechanisms of AA effect on performance, should we find one. Specifically, we added an informational prime similar to those that have been shown to activate a stereotype threat effect in previous studies (e.g., Spencer et al. 1999) with a paragraph in the description of the quota policy reading, *“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 each group of four (two men, two women) will get the bonus, as long as at least one of the two is a woman.* Hence, if AA does not act as a prime that triggers the stereotype threat effect, or if it does act as such but only partially, we would expect performance under AA-I, the AA condition with information, to be lower than performance under AA-no-I, the AA condition without the additional information.

The third round was identical to the first round: participants were given piece-rate incentives according to their scores. We use the third round to judge ceiling effect in performance and the effect of fatigue.⁷ While participants were given incentives to perform in all three rounds, they had the greatest incentives to perform in round 2, when they also had a chance to win the extra

⁷ Ceiling effect refers to the possibility that subject reach their maximum performance in round 1 and therefore could not increase their performance in round 2. Fatigue refers to the possibility that performance in round 2 is lower than round 1 due to exhaustion. If the performance in round 3 is found to be higher this rules out both explanations.

bonus. Hence, if the marginal cost of effort equals the marginal incentive, we would expect weakly higher effort in the second round compared to the other two rounds.

After completing the three rounds, but before learning about earnings and awards of the bonus, participants filled out a questionnaire in which they reported their SAT scores (quantitative and verbal), the extent to which they exerted effort on our exam, two Cognitive Reflection Test (CRT) questions used to proxy IQ,⁸ their expected score in round 2, and what they thought their chance of winning the bonus was. They were then informed of their earnings in the three rounds, told whether they had won the \$10 bonus, and paid privately in cash. See the Appendix for screen shots of the study. Average earnings from performance in the study (all rounds) not including the bonus were \$20.19; including the bonus, average earnings were \$25.43. That is, the size of the bonus was about half the average piece-rate earnings.

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-no-I condition, and 84 subjects in the AA-I condition—and each condition contained equal numbers of men and women.⁹ All subjects were undergraduate or graduate students of Harvard University recruited from the lab’s subject pool and were 20 years old on average. The self-reported average quantitative SAT score was 729.73 and the average verbal SAT score was 719.46. Balancing tests indicate that the randomization was successful and treatment samples within each condition are similar across key observables such as age and previous test scores (see Table 1).

⁸ The two questions were as follows: (1) “A bat and a ball cost \$1.10 in total. The bat costs a \$1 more than the ball. How much does the ball cost?” and (2) “If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?” These are two of three questions of the CRT that are thought to be associated with IQ (Frederick 2005).

⁹ Two subjects were dropped from our analysis because they left before the end of the experiment. Five additional male subjects who had missing survey questions about their SAT performance were also excluded from analyses where we control for measures of ability.

3. Results

To investigate the effect of gender-based AA policies on performance in a competitive math test, we focus on the effect of AA on the change in test scores between the first and second rounds.¹⁰ We use a weighted least squares (WLS) model that adjusts for a systematic relationship between variance in score and number of questions attempted.¹¹ We also explore the effect of AA on the individual's self-reported measure of effort and on his or her predicted score, which may capture changes in effort.

3.1. The Net Effect of Affirmative Action

Table 2 presents descriptive information on gender differences in performance for each round. In all rounds, men's scores are, on average, higher than women's: in round 1, men's average score is 6.36 and women's is 5.65; in round 2, men's average score is 7.31 and women's is 6.45; and in round 3, men's average score is 7.54 while women's is 7.06. Examining the average number of questions attempted, we find systematic gender differences that appear to reflect different response strategies: men answered more questions in every round—8.91 questions on average in round 1, 11.51 in round 2, and 12.20 in round 3, whereas women answered 8.32 questions on average in round 1, 9.89 in round 2, and 11.26 in round 3. We find that these differences in response strategy are statistically significant in round 2 and 3. That is, under piece-rate payment schedule and before the introduction of competition, there is no gender difference in response to incentives. Gender differences in response strategies on competitive tests such as

¹⁰ Test scores determine payments and capture both the quantity and the accuracy of the responses.

¹¹ A technical concern of using ordinary least squares (OLS) is heteroskedasticity, in which 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 pattern of results whether we use WLS, OLS, or bootstrap regression models. Also note that using number of questions attempted in round 2 as weights rather than the difference in questions attempted in round 2 minus round 1 is appropriate since a similar regression explaining the score in round 2 is equivalent to the regression explaining the difference in scores between round 2 and round 1. The only change would be the coefficient on the right hand side of the round 1 score.

the SAT and GRE are well documented in experimental studies by Baldiga (2013), Ben-Shakhar & Sinai (1991), and Hirschfeld et al. (1995).

The observed gender differences in baseline ability and response to competition in round 2 suggest that the impact of AA may differ for men and women. Guided by these differences, we separately analyze the effect of AA on performance progress for each gender. We also control for ability (using the first-round scores), response strategy (using the number of questions attempted), age, and self-reported SAT scores.¹²

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 adds an interaction term of the AA dummy variable with the first-round score; the latter allows for a heterogeneous effect of AA by ability. That effect could have several causes, such as low-ability men giving up under AA or low-ability women getting the greatest boost because of their increased opportunity to win.

Table 3 reports the results of pooling both AA conditions together.¹³ We find that, regardless of the specification considered, baseline ability (first-round score) has a significant negative effect on the change in scores between the first two rounds.¹⁴ This reflects the fact that the higher the baseline score, the more difficult it is to improve. We also find that the self-reported quantitative SAT score has a positive significant association with the change in scores; however, the self-reported verbal SAT score does not. The number of questions attempted in the first round, which may capture response strategy,¹⁵ is associated with a significantly higher performance progress

¹² There is no significant gender difference in average quantitative SAT score in our sample: women's score is 736 while men's is 718 (one-sided t -test yields $p = 0.11$).

¹³ The results in the table exclude five subjects missing data on SAT performance but the results are robust to an alternate specification that includes these subjects and does not control for SAT performance.

¹⁴ The dependent variable is the change in score from round 1 to round 2. Whether we use the change in score or round 2's performance level does not change the result. This is because round 1 score is our proxy of ability and we control for it in the regression. For ease of interpretation—improvement or decline in performance—we opted to use performance difference rather than level.

¹⁵ Indeed, in a regression of the success rate in round 2 on the number of attempts, gender, and the interaction of the two, 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 -0.012 , $p = 0.107$). Nevertheless, the gender effect is positive up to 14 attempts,

for women (effect of 0.442 or 0.471, depending on the specification, both with p -value < 0.001), but an insignificant effect for men.

Turning to the main question—the effect of AA on performance progress—we find that, regardless of ability (Table 3, column 1), AA has a positive, albeit insignificant, effect on women. Allowing for the possibility that high- and low-performing women are affected differently (Table 3, column 2), we find a positive significant main effect of AA (1.889, p -value = 0.019) that declines with ability (-0.23 ; p -value = 0.035). For the average woman in the sample, the overall effect of AA is positive and accounts for about 10% of the score in round 1. That is, the average woman's score progress is better under AA than without it. However, the overall effect of AA becomes negative for women whose first-round score is over 8.21—approximately the top 20% of female performers.^{16,17,18} Note that this result is not due to regression to the mean. While regression to the mean may exist in this setting—that is, the performance of high-ability women declines in the second round while that of the low-ability women increases—we compare women of similar ability level across AA conditions. We find that the effect of AA on high ability women is negative, meaning that even if high ability women reduce performance from round 1 to round 2 due to regression to the mean, the reduction in performance is stronger with AA than without it. For men, we find an insignificant positive effect of AA regardless of the specification. That is, the main effect of AA on men's performance progress, regardless of ability (Table 3, column 3), is 0.566 and is insignificant. When AA is allowed to have a differential effect by ability (Table 3,

representing about 92% of the women in our sample. This is consistent with a different response strategy across gender, in which men attempt to answer more questions at the cost of accuracy.

¹⁶ Looking at percentiles, approximately 80% of women have a first-round score lower than 8 in round 1, which is very close to the 8.21 cutoff.

¹⁷ 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.

¹⁸ Examining the correlation between the score in round 1 and the score in round 2, as well as between the number of questions attempted across the two rounds, we find very high correlation of about 0.8 for both. That is, the effect is not driven by high-ability women who performed poorly in the first stage.

column 4), its main effect is positive (1.051) and its interaction with ability is negative (-0.067), both are insignificant.

To test whether the overall negative effect on the high-ability women subgroup is in fact significant, we use the bootstrap method that allows such testing in spite of sample size issues and does not require any parametric assumptions. Specifically, we split the women in our sample into three groups according to their first-round scores: low ability (scores below 5; bottom 50%), mid ability (scores between 5 and 8; between 50% and 80%), and high ability (scores above 8; top 20%). This split is guided by the WLS regression result: the top 20% corresponds closely to the group for whom the overall effect of AA appears to be negative.¹⁹ We then calculate, for each ability subgroup, the difference in (mean) scores between round 2 and round 1. We do so separately for women under AA and women 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 ability subgroup with repetition, resulting in a distribution of diff-in-diff (see Figure 1). This allows us to ask whether the average diff-in-diff is negative and significant for the subgroup of high-ability women. We find that the mean diff-in-diff for the high-ability women is negative in 93% of the random sample, with its mean equal to -1.36 and significantly different from zero. That is, the bootstrap exercise demonstrates that the total effect of AA is negative for high ability women.

We further complement the bootstrap exercise with a regression analysis using the division of women into the three ability groups described above coded as an ordered group dummy (0,1,2) for low-, mid-, and high-ability women; this exercise allows for a nonlinear relationship between ability and the effect of AA.²⁰ The results (see Table 4) are consistent with those obtained from the

¹⁹ The top-ability group definition is guided by the regression results suggesting to focus on the top 20% of female participants. The additional split into low- and mid-ability groups was done based on the premise that below-the-median performance is considered low, and then it follows that performance above the median but not at the top 20% is average.

²⁰ Using 40% and 60% cutoffs to determine low- and mid-ability groups (i.e., 40% cutoff means that below 40% is considered low, and performance in the 40-80% range is mid-level ability) do not change the qualitative results. Separately, if one were to use the OLS results the suggested round 1's score cutoff for the negative effect of AA is slightly lower at about 7.1. Using this cutoff to define the high ability group, the results are robust

bootstrap approach: the effect of AA on low-ability women is positive (1.43) and at least marginally significant (one-sided t -test for a positive effect yields $p = 0.03$. The two-sided t -test yields $p = 0.056$); its effect on mid-ability women is overall negative (-0.118) but insignificant, and its effect on high-ability women is negative (-1.67) and at least marginally significant (one-sided t -test for a negative effect yields $p = 0.03$. The two-sided t -test yields $p = 0.067$.)²¹ For robustness check, we repeat this exercise with two additional cutoffs—top 25% and top 15%—for defining the high ability group. The results are robust for these different cutoffs—see Table A1 in the Appendix.

Before moving on to explore possible mechanisms that could drive this heterogenous effect of AA on performance, we may wonder whether the result is robust to using a different measure of ability. To explore this issue, we repeat the exercise using participants' round 3 scores as our measure of ability. We find that 22 out of the 25 top female performers based on round 1 performance are also classified as top performers based on their round 3 score, which is reassuring and suggests that the ability classification based on round 1's score is not due to luck. Using this alternative measure (based on round 3 scores) gives rise to a similar pattern, with a negative coefficient estimated for the interaction of affirmative action and the high ability group indicator. The coefficients are of similar magnitude to those in the main analysis, however the statistical significance of the interaction is weaker with p -values at $p=0.105$ when using a separate dummy for mid- and high-ability groups and $p=0.12$ when using an ordered ability-group dummy. This is to be expected because individuals were already exposed to the treatment and round 3 score is therefore a noisier measure of ability.²²

²¹ Coding the score group dummy as an ordered variable (0,1,2) allows for nonlinear relationship between ability and performance, however it assumes a linear effect moving from one group to another. Using dummies separately for the medium-ability subgroup and for the high-ability subgroup, we find again that AA has an insignificant negative effect on the medium-ability group and a significant negative effect on high-ability women.

²² We cannot construct meaningful ability groups using the self-reported SAT quantitative score or the CRT questions because there is not enough variation in these variables among participants in our experiment. Conceptually, we also believe that using the score in round 1 is a more appropriate proxy for ability because it is observed rather than self-reported, it was measured at the time the subjects took the test, and it is free of any treatment effects

3.2. The Underlying Mechanisms

While the result highlighting the possibility of an unintended negative effect of AA on performance is important in and of itself, an interesting issue for policymakers is the possible underlying mechanisms driving this result. AA might be affecting performance in several ways: (1) by affecting an incentive to exert effort; (2) by acting as a negative prime triggering the stereotype threat effect; and (3) by changing the perceived sex composition of the tournament, leading women to focus on a single- rather than mixed-sex competition. It may also affect performance through a combination of all or some of these mechanisms. Below we review each of these potential mechanisms and their expected effects.

The first mechanism, incentive to change effort, refers to the fact that AA affects women's probability of winning the bonus, which, in turn, changes the incentive to exert effort. Without AA, a woman must outperform two of three competitors to win the bonus. She also knows that the score of one of her competitors is drawn from the same (female) score distribution as hers and that the other two competitors' scores are drawn from the (male) distribution with a higher mean. With AA, her chance of winning the bonus increases because it is now sufficient to outperform just one other competitor, and she also knows that this other competitor is a woman whose ability is drawn from the female performance distribution.

The second mechanism is the stereotype threat effect, whereby reminding individuals that they belong to a group that stereotypically performs worse than other groups adversely affects their performance on the given task.²³ In our context, AA may be viewed as a prime triggering the stereotype threat effect and may therefore reduce women's performance.

²³ This second potential mechanism—the stereotype threat effect—was introduced by the psychology literature. It suggests that reminding individuals that they belong to a group that stereotypically performs worse than other groups on a task adversely affects their performance on that task, especially if the task is challenging. For example, studies document that reminding African-American students of their race leads them to perform significantly worse on verbal 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). Although most studies in the stereotype threat literature find a stereotype threat effect, others find opposite (Wei 2009, 2012) or no effects (Stricker 1998, Fryer et al. 2008). This has also recently been shown by Iriberry and Rey-

The third mechanism is the single-sex competition, and it refers to the findings that women are less effective in mixed-sex competitive environments. If it leads women to perceive the competition as a single-sex competition, AA may circumvent this effect of mixed-sex competition and help women improve their performance.

Of the three potential mechanisms that we have identified, we can directly test the effect of single-sex competition by running an additional condition in which the second-round competition is between two women.²⁴ Thirty-four women from the same subject pool participated in this condition.²⁵ The setting for this women-only condition was identical to that of the other experimental conditions described above. That is, participants were asked to solve GRE quantitative questions in three rounds. In round 1 and round 3, they were given a piece-rate payment; in round 2, however, they were assigned to single-sex groups of two women, where the person to win the bonus was the woman with the highest score.

We then test whether, in round 2, women's performance in the single-sex condition differs from women's performance in the control group, where subjects were assigned to mixed-sex groups and were not subject to AA. Table 5 shows no significant difference between the two conditions: neither the main effect of women-only competition 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.

Turning to the remaining mechanisms—incentive and stereotype threat effect—only the incentive effect can account for the positive effect of AA on low-ability women. While this result

Biel (2017), noting that gender stereotype threat effect is most likely to present itself in tasks that are perceived as unfavorable to women and where the presence of rivals is salient.

²⁴ We note that the incentives in a single-sex competition and mixed-sex AA competition are not exactly comparable because they are dependent on ability level. That is, in the mixed-sex AA competition a woman can win the bonus if she is better than the other woman, and if she is not better than the other woman, she could still win the bonus if she is better than the other two men. However, importantly, if AA change women's mindset to think of their competition as a single-sex competition then this treatment should capture the incentives associated with that view.

²⁵ We ran women-only conditions during the same sessions as all other conditions, so the observable gender composition was mixed during all sessions.

affirms the existence of an incentive effect, it does not rule out the stereotype threat effect, which may still exist even if dominated by incentive. For the high-ability women, we find a negative AA effect on performance. The incentive mechanism for these women can go both ways because, for high-ability women, who know they are of high ability, the chances of winning under AA are high even without their putting in much effort. Hence, this mechanism could lead high-ability women to exert more, the same, or less effort under AA.²⁶ The overall negative effect of AA on the performance of high-ability women can therefore be a result of either a negative incentive effect (with or without the existence of the stereotype threat effect) or a positive incentive effect, similar to that observed for low-ability women, which is dominated by a negative stereotype threat effect.

To look into stereotype threat effect, we first confirmed that Harvard students believe women underperform compared to men on the GRE math questions using a survey.²⁷ Next we use the two AA conditions with and without an informational prime (AA-I, and AA-no-I, respectively)—where the information provided in the AA-I condition is similar to the informational prime used in the psychology literature to trigger stereotype threat effect (see e.g., Spencer et al. 1999)—to test for differential effect. If AA does not act as a prime triggering a

²⁶ To illustrate why these different effects are possible, we can think of each participant's problem in round 1 as choosing an effort level to maximize $Reward(e) - c(e)$, where $Reward(e)$ is the piece rate earning as a function of effort e and $c(e)$ is the cost of exerting effort level e and of the problem in round 2 as choosing e to maximize $p(e;AA)Bonus + Reward(e) - c(e)$, where $p(e;AA)$ is the probability of winning the bonus with effort level e under a specific AA condition. In this toy model, the optimal effort level in round 1 equates the marginal benefit of higher effort in piece rate earnings with the marginal cost of effort. In round 2, the marginal benefit is again the increased piece rate earnings *and* an elevated probability of earning the bonus. Hence, unless the cost function changes with AA—which is a way of thinking of stereotype threat—it would generally be optimal for women to increase effort in round 2. However, it is possible that high-ability women might choose to reduce effort (or keep it constant) compared to when in a mixed-sex competition without AA if they are confident that under AA they will win the bonus. In that case, the marginal increase in probability to win the bonus with effort may be lower under AA.

²⁷ 75 Harvard students participated in a survey regarding the gender gap in GRE math performance. Each subject was asked to report his or her belief about women's GRE math scores compared to men's—specifically, whether women's scores are much lower (1), lower (2), the same (3), higher (4), or much higher (5) than men's scores. Overall, the Harvard students believe that women's GRE quantitative performance is lower than men's. This holds true whether they are asked about the general population that takes the GRE (average response is 2.7, significantly different from 3 at the 1% level) or about the Harvard student body (average response is 2.78, significantly different from 3 at the 1% level).

stereotype threat effect (or does so only partially), and assuming that the informational prime does, we would expect performance to be lower under AA-I. To test whether there is a differential effect of AA with and without an informational prime, we run an analysis similar to that shown in Table 4 but with an additional indicator variable “Info,” which takes the value of 1 if a participant was assigned to the AA-I treatment and of zero otherwise.²⁸ We also interact this indicator variable with our measure of ability and find that both its main effect and its interaction with the ability group are insignificant (see Table 6). There are two possible explanations consistent with this result: either that AA fully act as a prime triggering stereotype threat or that stereotype threat effect does not exist in this environment.

Because the experimental design cannot disentangle these two mechanisms, we look at our survey measures to examine whether high-ability women’s effort changes in response to AA. We start by investigating participants’ self-reported effort exerted during the study, as indicated in an exit questionnaire (responses range from 1 to 7, where 7 represents the highest effort). Table 7 reports the results of Ordered Probit regressions analyzing those responses using the same specification as in our main analysis (Table 3): AA dummy variable, first-round score, and the interaction of the two, plus controls that include the number of questions attempted in the first round, age, and self-reported SAT scores. Two specifications were considered as before: one using the continuous measure of ability, forcing a linear relationship, and the other using ability groups, allowing for a nonlinear relationship. Two women reported an unusually low level of effort of less than or equal to 2 (the median effort is 6); we 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 report exerting 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. Taking into account all the women, including the two outliers, the effect is always positive albeit sometimes insignificant. The results, therefore, suggest that women’s efforts do not decrease in response to AA across the board.

²⁸ The information prime describes women’s inferior performance relative to men on the GRE, which is, on average, 15% lower. See Section 2 for details.

The second measure that we examine is the participants' round 2 score predictions, submitted before any feedback on actual performance was given. If high-ability women exerted less effort under AA, the expectation is that they would predict lower round 2 scores than participants in the No AA condition. Table 8 shows that, overall, women under AA reported higher scores in round 2 (average predicted score is 6.2) than those in the control (with average predicted score of 5.8). While this positive difference is not statistically significant for low- and mid-ability women, it is significant at the 5% level for high-ability women. This evidence is again consistent with women not decreasing their efforts in response to AA.

The two self-reported measures of effort suggest that women—in particular, high-ability women—did not reduce their effort in response to AA in our study. While this evidence is not conclusive, it is nevertheless suggestive of the coexistence of the incentive mechanism and the stereotype threat effect.

3.3. Response Strategy

Test scores are influenced by both the number of questions attempted and the success rate on these questions. That is, an increase in a given score can be obtained by increasing the number of questions answered while keeping the success rate fixed, by improving the success rate while the number of questions answered is fixed, or both.

Table 9 shows the regression analysis of the number of questions attempted in round 2 on a similar specification as in prior analyses: a dummy of AA, baseline ability (score in round 1), and their interaction. We also control for the number of questions attempted in round 1, age, and self-reported SAT scores. We find a positive but insignificant main effect of AA on the number of questions that women attempt and this effect does not vary with ability.

We then proceed to examine the effect of AA on success rate. Again, we use a regression analysis with the same specification as above, but with the success rate in round 2 as the dependent variable. Table 10 shows that the main effect of AA on success rate is positive and

significant and that this effect declines with ability. For the highest-ability women, those whose average score in round 1 is over 11, the overall effect of AA on success rate is negative.

4. Conclusions

This paper examines the effect of gender-based AA on quantitative GRE performance in an incentivized and competitive environment. While the findings show that AA positively affects the performance of most women, we find a surprising negative effect on the performance of women of the highest ability. This result has not been detected by the literature thus far, possibly owing to task differences. As we noted above, the GRE math task has several distinct properties that other tasks used in the literature do not share. First, participants are very familiar with this task, as they have all taken at least the SAT and therefore have a good understanding of their relative ranking in the performance distribution. Second, there is an established stereotype of gender gap in performance, which accurately reflects reality. Together, this knowledge and the gender gap in performance result in different incentives by ability. Furthermore, the established stereotype combined with the fact that the task is relatively challenging may lend itself to a stereotype threat effect.

We discuss three potential mechanisms that, either separately or together, may drive our results. Of these three potential mechanisms, we test the single-sex competition directly and find that this cannot explain the results. Our experiment cannot disentangle the remaining two mechanisms—the incentive effect and stereotype effect. Nevertheless, we use self-reported measures of effort and predicted performance to shed light on the effect of AA on effort exerted. We find no evidence that women of any ability reduced their effort in response to AA; yet given that these are self-reported measures, the results are only suggestive.

An important finding that can be drawn from our results is that when policymakers consider the impacts of introducing an AA policy, they should account for the possibility that this policy may have a negative impact on performance in the group of intended beneficiaries. When possible, policymakers might avoid this effect by not announcing the policy rule explicitly. Future

research disentangling the incentive and stereotype threat mechanisms is important to developing specific policies. If the effect is solely due to incentives, AA may not be the optimal policy to use. For example, if the objective is to select women of highest ability, their reduction in effort could lead them to fall behind. In that case, AA could be detrimental to overall human capital accumulation if high-ability women reduce effort in response to AA and AA is used in different stages of life. That is, if effort is strategically reduced in response to AA, it can lead women to underperform and fuel gender stereotypes in spite of women's true ability. On the other hand, if the effect is solely due to stereotype threat, then AA may be an effective policy if coupled with education at an early age.

As noted in the paper, participants in the study were students in a highly selective institution, and it may be argued that they are not representative of the effect of AA in the field. However, AA policies are often designed to motivate candidates of the highest abilities, so the results in this study are directly relevant to these contexts. Second, it is possible that the effect is driven by relative ranking within the group rather than by absolute skill. If so, the effect found in this study may also occur in other groups of lower abilities as long as the high-ability individuals are defined relative to the group they are competing against.

With the substantial interest around the world in AA policies aimed at advancing women, it is important to fully understand their potential effect. This paper provides the first direct experimental evidence that AA policies may have unintended negative consequences on performance.

References

- American Psychological Association. 2012. Brief of Experimental Psychologists et al. as Amici Curiae Supporting Respondents, Fisher v. University of Texas, (August 13, 2012) (No. 01-1015).
- Balafoutas, Loukas, and Matthias Sutter. 2012. Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science* 335:579–582.
- Baldiga, Katherine. 2013. Gender differences in willingness to guess. *Management Science* 60(2): 434-448.
- Ben-Shakhar, Gershon and Sinai, Yakov, 1991. Gender differences in multiple-choice tests: the role of differential guessing tendencies. *Journal of Educational Measurement*, 28(1), pp.23-35.
- 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.
- Frederick Shane. 2005. Cognitive Reflection and Decision Making. *Journal of Economic Perspective* 19(4): 25-42.
- Fryer, Roland, Steven Levitt, John List. 2008. Exploring the impact of financial incentives on stereotype threat. *American Economic Review: Papers and Proceedings* 98(2): 370-375.
- Gneezy, Uri, Muriel Niederle, and Aldo Rustichini. 2003. Performance in competitive environments: Gender differences. *Quarterly Journal of Economics* 118: 1049–1074.
- Hirschfeld, Mary, Moore, Robert L. and Eleanor Brown. 1995. Exploring the gender gap on the GRE subject test in economics. *The Journal of Economic Education*, 26(1): 3-15.
- Hoff, Karla, and Priyanka Pandey. 2006. Discrimination, Social Identity, and Durable Inequalities. *The American Economic Review Papers and Proceedings*, 96(2): 206-211.

- Imbens, Guido, Donald B Rubin, Gary King, Richard A Berk, Daniel E Ho, Kevin M Quinn, James D Greiner, Ian Ayres, Richard Brooks, Paul Oyer, and Richard Lempert. 2012. "[Brief of Empirical Scholars as Amici Curiae.](#)" Filed with the Supreme Court of the United States in Abigail Noel Fisher v. University of Texas at Austin, et al.
- Iriberry, Nagore, and Pedro Rey-Biel. 2017. Stereotypes are only a threat when beliefs are reinforced: On the sensitivity of gender differences in performance under competition to information provision. *Journal of Economic Behavior & Organization* 135: 99-111.
- Leslie, L.M., Mayer, D.M. and Kravitz, D.A., 2014. The stigma of affirmative action: a stereotyping-based theory and meta-analytic test of the consequences for performance. *Academy of Management Journal*, 57(4), pp.964-989.
- 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 Year 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.
- Sander, Richard. H., 2010. Listening to the Debate on Reforming Law School Admissions Preferences. *Denver University Law Review* 88: 889.
- Schuck, Peter H., 2002. Affirmative action: Past, present, and future. *Yale Law & Policy Review*, 20(1): 1-96.
- 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., 1997. A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist* 52(6):613.

- 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. In *Advances in experimental social psychology* (Vol. 34, pp. 379-440). Academic Press.
- 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.
- Willingham, W. W., & Cole, N. S. (1997). Gender and fair assessment. Mahwah, NJ: Lawrence Erlbaum.

**Table 1: Summary Statistics: Independent Variables
Control vs. AA**

	Control (mean)	AA (mean)	P-Value
Age	20.22	20.07	0.55
SAT Quantity	722.66	734.13	0.45
SAT Verbal	717.09	722.81	0.73
CRT Questions	1.05	1.12	0.53
N	79	167	

**Table 2: Summary Statistics: Dependent Variables
By Gender and Round**

		Male (mean)	Female (mean)	P-Value
Round 1	Score	6.36	5.65	0.156
	# Questions	8.91	8.32	0.201
	Ratio correct	0.76	0.72	0.236
Round 2	Score	7.31	6.45	0.107
	# Questions	11.50	9.89	0.001
	Ratio correct	0.69	0.71	0.386
Round 3	Score	7.54	7.06	0.397
	# Questions	12.20	11.26	0.037
	Ratio correct	0.68	0.68	0.961
N		123	123	

Table 3: The Effect of Affirmative Action on Performance Progress

	(1) Female	(2) Female	(3) Male	(4) Male
Affirmative Action (AA)	0.436 (0.432)	1.889** (0.795)	0.566 (0.532)	1.051 (0.904)
Score in 1 st Round (Score R1)	-0.723*** (0.078)	-0.572*** (0.093)	-0.372*** (0.141)	-0.340** (0.137)
(AA)x(Score R1)		-0.230** (0.108)		-0.067 (0.128)
# Questions in 1 st Round	0.442*** (0.069)	0.471*** (0.075)	0.187 (0.132)	0.171 (0.130)
Age	0.000 (0.106)	-0.052 (0.111)	-0.173 (0.180)	-0.160 (0.177)
SAT Quantitative	0.018*** (0.003)	0.017*** (0.003)	0.022*** (0.006)	0.023*** (0.006)
SAT Verbal	0.002 (0.003)	0.002 (0.003)	0.002 (0.004)	0.002 (0.004)
Constant	-13.242*** (3.974)	-13.153*** (3.931)	-13.356** (5.905)	-14.058** (6.019)
N	123	123	118	118
Adjusted R ²	0.359	0.380	0.176	0.171

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% level; ** indicates significance at the 5% level; *** indicates significance at the 1% level. *t* indicates $p < 0.15$.

Table 4: The Effect of Affirmative Action on Performance, by Group

	(1) Female	(2) Male
Affirmative Action (AA)	1.434* (0.744)	0.979 (0.705)
1 st Round Score Group	-1.071* (0.582)	-1.270* (0.674)
AA x 1 st Round Score Group	-1.552** (0.670)	0.108 (0.637)
# Questions in 1 st Round	0.195** (0.095)	0.013 (0.082)
Age	-0.018 (0.140)	-0.123 (0.177)
SAT Quantitative	0.012*** (0.004)	0.022*** (0.006)
SAT Verbal	-0.001 (0.004)	0.001 (0.004)
Constant	-7.887* (4.578)	-13.626** (6.313)
N	123	118
r ²	0.229	0.203
T_Low_pos	0.028	0.084
T_High_neg	0.034	0.901

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 affected by the AA. T-High(-) displays the result of a *t*-test calculating the probability that an individual in the higher performance group was negatively affected by the AA. Robust standard errors are reported in parentheses. * indicates significance at the 10% level; ** indicates significance at the 5% level; *** indicates significance at the 1% level. *t* indicates $p < 0.15$.

Table 5: Women-Only Competition

	(1)	(2)
Women only	0.235 (0.802)	0.366 (0.743)
Score in 1 st Round (Score R1)	-0.451*** (0.124)	
1 st Round Score Group		-0.667 (0.586)
(Women only)x(Score R1)	0.025 (0.095)	
(Women only)*(1 st Round Score Group)		-0.248 (0.592)
# Questions in 1 st round	0.311** (0.122)	0.074 (0.078)
Age	-0.066 (0.120)	-0.082 (0.146)
SAT Quantitative	0.012* (0.006)	0.004 (0.007)
SAT Verbal	0.003 (0.003)	0.001 (0.004)
N	72	72
r ²	0.193	0.057

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. t indicates $p < 0.15$.

Table 6: Affirmative Action and Informational Prime

	(1) Female	(2) Male
Affirmative Action (AA)	1.696** (0.755)	1.350 ^t (0.830)
Info	-0.519 (0.701)	-0.616 (0.958)
1 st Round Score Group	-1.071* (0.586)	-1.277* (0.683)
AA x 1 st Round Score Group	-1.601** (0.741)	-0.048 (0.784)
Info x 1 st Round Score Group	0.064 (0.753)	0.257 (0.842)
# Questions in 1 st Round	0.196** (0.096)	0.021 (0.082)
Age	-0.026 (0.135)	-0.135 (0.181)
SAT Quantitative	0.012*** (0.004)	0.021*** (0.006)
SAT Verbal	-0.001 (0.004)	0.001 (0.004)
Constant	-7.644* (4.526)	-12.898* (6.811)
N	123	118
r ²	0.234	0.207

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. t indicates $p < 0.15$.

Table 7: The Effect of Affirmative Action on Effort, Women

	(1)	(2)	(3)	(4)	(5)	(6)
Affirmative Action (AA)	0.405* (0.211)	0.473** (0.221)	0.549 ^t (0.374)	0.878** (0.404)	0.497 ^t (0.318)	0.668* (0.353)
Score in 1 st Round (Score R1)	0.036 (0.049)	0.041 (0.049)	0.052 (0.052)	0.086 ^t (0.057)		
(AA)x(Score R1)			-0.026 (0.051)	-0.071 (0.055)		
1 st Round Score Group					0.187 (0.247)	0.392 ^t (0.270)
AA x 1 st Round Score Group					-0.093 (0.256)	-0.200 (0.279)
# Questions in 1 st round	-0.017 (0.048)	-0.005 (0.046)	-0.014 (0.049)	0.006 (0.045)	-0.007 (0.037)	-0.002 (0.037)
Age	0.017 (0.064)	0.009 (0.068)	0.012 (0.067)	-0.006 (0.073)	0.012 (0.066)	-0.001 (0.071)
SAT Quantitative	0.005** (0.002)	0.002 (0.003)	0.005** (0.002)	0.002 (0.003)	0.005** (0.002)	0.002 (0.002)
SAT Verbal	-0.002 (0.002)	-0.003 (0.002)	-0.002 (0.002)	-0.003 (0.002)	-0.002 (0.002)	-0.003 ^t (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. t indicates $p < 0.15$.

Table 8: Women self-reported predicted round 2 scores by ability and AA condition (prior to learning actual score).

Ability	Low-Ability	Mid-Ability	High-Ability	Overall
AA-no-I & AA-I	4.5	6.17	10.36	6.2
No AA	3.97	6.35	8.11	5.8
Diff	0.53	-0.18	2.25**	0.4

* indicates significance at the 10-percent level; ** indicates significance at the 5-percent level; *** indicates significance at the 1-percent level.

Table 9: Number of Questions

	(1) Female	(2) Female
Affirmative Action (AA)	-0.529 (0.904)	-0.085 (0.673)
Score in 1 st round (Score R1)	0.178 (0.149)	
(AA)x(Score R1)	-0.037 (0.138)	
1 st Round Score Group		1.252* (0.654)
AA x 1 st Round Score Group		-0.570 (0.711)
# Questions in 1 st Round	0.526*** (0.197)	0.545*** (0.148)
Age	0.091 (0.249)	0.074 (0.244)
SAT Quantitative	0.009* (0.005)	0.008** (0.004)
SAT Verbal	-0.005 ^t (0.004)	-0.006 ^t (0.004)
N	123	123
r ²	0.600	0.612

Notes: WLS regressions. Dependent variable: 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. t indicates p<0.15.

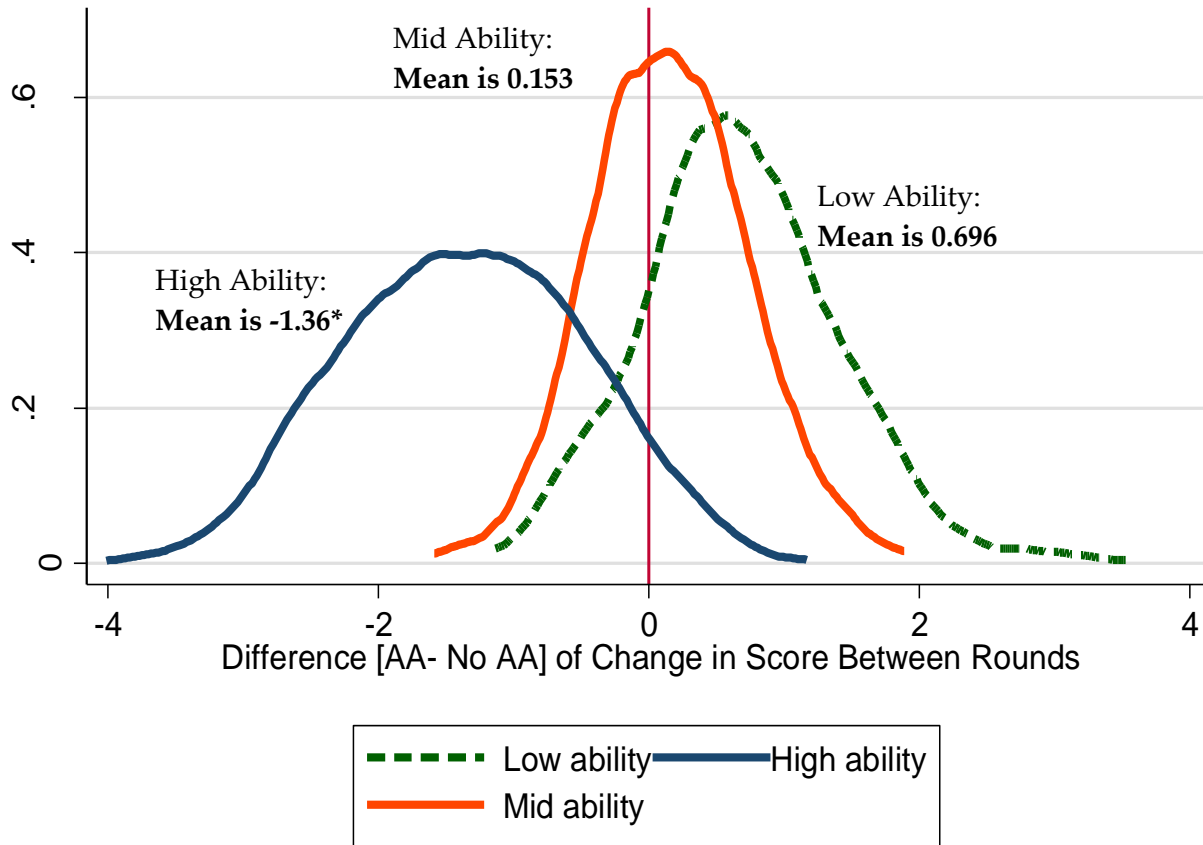
Table 10: Success Rate

	(1) Female	(2) Female
Affirmative Action (AA)	0.149** (0.064)	0.120** (0.053)
Score in 1 st Round (Score R1)	0.019** (0.009)	
(AA)x(Score R1)	-0.013* (0.008)	
1 st Round Score Group		0.064 ^t (0.040)
AA x 1 st Round Score Group		-0.051 (0.043)
# Questions in 1 st Round	-0.001 (0.009)	0.002 (0.006)
Age	0.004 (0.010)	0.004 (0.010)
SAT Quantitative	0.001*** (0.000)	0.002*** (0.000)
SAT Verbal	0.001** (0.000)	0.001** (0.000)
N	123	123
r ²	0.416	0.409

Notes: OLS regressions. Dependent variable: 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. t indicates p<0.15.

Figure 1

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



Appendix:

Table A1: Robustness Checks²⁹
Different Cutoffs for Being Considered a Top Performer, Females

	(1)	(2)
Top Cutoff:	75%	85%
Affirmative Action (AA)	1.427* (0.756)	1.132t (0.739)
1 st Round Score Group	-1.040* (0.559)	-1.930*** (0.522)
(AA)x(1 st Round Score Group)	-1.363** (0.637)	-1.238** (0.618)
# Questions in 1 st Round	0.174* (0.097)	0.245*** (0.084)
Age	-0.032 (0.14)	0.014 (0.125)
SAT Quantitative	0.012*** (0.004)	0.014*** (0.004)
SAT Verbal	-0.002 (0.004)	0.002 (0.004)
Constant	-6.580t (4.483)	-11.416** (4.448)
N	123.000	123.000
r ²	0.234	0.30
T_Low_pos	0.031	0.064
T_High_neg	0.061	0.043

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 affected by the AA. T-High(-) displays the result of a *t*-test calculating the probability that an individual in the higher performance group was negatively affected by the AA. Robust standard errors are reported in parentheses. * indicates significance at the 10% level; ** indicates significance at the 5% level; *** indicates significance at the 1% level. *t* indicates $p < 0.15$.

²⁹ The score group variables is an ordered dummy variable (0,1,2) for the low-, mid-, and high-ability women group. The results are similar using separate dummy variables for each ability group.

Screen shots for the AA-I condition.

In this experiment you will be asked to solve math problems taken from the GRE quantitative section.

The questions will appear on the screen, one at a time, together with 5 possible answers. Each question has only one correct answer, and your task is to select the correct answer for each question.

You are allowed to use only the paper and pencil provided. Calculators, handheld computers, cellphones, friends, or any other helping device is not allowed.

Click on the 'OK' button to see an example of a problem.

OK

What is $2+3$ equal to?

This is just a practice trial so that you are familiar with the layout of the questions. The actual questions will be more difficult than this example. To select one of the answers click on it with the mouse. If you change your mind, you can click on a different answer to change your selection. Once you have made up your mind, click on the OK button to continue (Please select the correct answer and continue to the next screen).

- 1
- 2
- 3
- 4
- 5

OK

In the actual task, after you solve one of the questions, you will be given another like it.

Here are further details regarding the experiment:

Time:

You will have 10 minutes to solve as many math problems as you can.

Payment:

You will earn \$1 for each correctly solved question, but you will lose \$0.25 for each incorrectly solved question (so nothing can be gained from guessing).

If you have any questions, please press the assistance request button located in front of you (below the screen). Otherwise click OK to continue.

OK

You will now begin the math test.

Remember:

You have **10 minutes** to solve as many math problems as you can.

Payment: \$1 per correctly solved question, minus \$0.25 per incorrectly solved question.

When you are ready, please press "Start".

Start

Please note

In the program we use the following symbols:

$6^2=6^2$ $6^{0.5}=\sqrt{6}$ $\text{Sqrt}(2)=\sqrt{2}$	That is, the symbol “^” means to the power of. That is, the symbol “sqrt” means square root.
--	---

$6/2 = \frac{6}{2}$	That is, “/” symbolizes division.
---------------------	-----------------------------------

$6 \leq 7$ means $6 \leq 7$	That is, the symbol “<or=” means “≤”.
-----------------------------	---------------------------------------

$\text{Pi}=\pi$	That is, the symbol “Pi” means “π”.
-----------------	-------------------------------------

OK

ROUND 1—10 MINUTES. Example:

The distance from point X to point Y is 20 miles, and the distance from point X to point Z is 12 miles. If d is the distance, in miles, between points Y and Z, then the range of possible values for d is indicated by

- $8 \leq d \leq 20$
- $8 \leq d \leq 32$
- $12 \leq d \leq 20$
- $12 \leq d \leq 32$
- $20 \leq d \leq 32$

OK

This 10 minutes are over.
Please click on the "OK" button to continue.

OK

Round 2

Now we start another round of 10 minutes, similar to the previous one.
In round 2 we divide all participants into groups.

Details on time, groups, and payment will be given shortly.

OK

Before we proceed, please answer the following:

What is your gender?

Female

Male

Please wait Patiently.

Once everyone is on the same page, we will give a code to continue.

Here are further details regarding this round:

Time:

You will have 10 minutes to solve as many math problems as you can.

Group:

You are in a group of 4, with two men and two women in the group.

Payment:

You will earn \$1 for each correctly solved question, but you will lose \$0.25 for each incorrectly solved question (so nothing can be gained from guessing).

In addition, 2 people in each group will earn a bonus of \$10 based on their performance. The exact rule which determines who gets the bonus will be described next.

If you have any questions, please press the assistance request button located below the screen. Otherwise click OK to continue.

OK

Bonus Payment for Contest Winners

As mentioned, in this round, on top of the payment for each correctly solved question, you could get a bonus of \$10. This depends on your performance, as explained next.

OK

Bonus Payment for Contest Winners

As mentioned, in this round, on top of the payment for each correctly solved question, you could get a bonus of \$10. This depends on your performance, as explained next.

How is performance determined?

For each participant we will calculate a **round 2 score**, which is calculated as follows:

$$\text{Round 2 Score} = \begin{array}{c} \text{(number of correctly solved questions in round 2)} \\ \text{minus} \\ 0.25 \times \text{(number of incorrectly solved questions in round 2)} \end{array}$$

For example:

If in round 2 you correctly solve 10 questions and incorrectly solve 4 questions, your score would be 9. Here is why: $10 - (0.25 \times 4) = 9$.

Note:

The penalty for incorrectly solved questions is designed such that on average there is no way to gain by guessing.

OK

Bonus Payment for Contest Winners

Who gets the bonus?

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.

In the case of a tie, the participant who spent the least time on the questions they correctly solved will get the bonus.





OK

In the next section you will be asked a series of questions to make sure you understand the winner selection process for the \$10.

Each time you will see scores from a hypothetical competition. You will see the scores of all 4 participants, and your task will be to indicate which two participants would win the \$10 bonus, if the scores were as displayed.





OK

Please indicate which two participants would win the \$10, if their scores were as the ones shown above each participant. To select one of the participants, click on the box below the figure representing them. If you have selected one of them and want to change that selection, just click on the selected box to unselect it. Click on the 'Submit' button once you have chosen the two participants.

5	2	6	4
			
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>





Submit

Please indicate which two participants would win the \$10, if their scores were as the ones shown above each participant. To select one of the participants, click on the box below the figure representing them. If you have selected one of them and want to change that selection, just click on the selected box to unselect it. Click on the 'Submit' button once you have chosen the two participants.

7	5	3	4
			
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>





Submit

Please indicate which two participants would win the \$10, if their scores were as the ones shown above each participant. To select one of the participants, click on the box below the figure representing them. If you have selected one of them and want to change that selection, just click on the selected box to unselect it. Click on the 'Submit' button once you have chosen the two participants.

4	2	5	8
			
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>





Submit

Please indicate which two participants would win the \$10, if their scores were as the ones shown above each participant. To select one of the participants, click on the box below the figure representing them. If you have selected one of them and want to change that selection, just click on the selected box to unselect it. Click on the 'Submit' button once you have chosen the two participants.

6	7	10	4
			
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Submit

Please indicate which two participants would win the \$10, if their scores were as the ones shown above each participant. To select one of the participants, click on the box below the figure representing them. If you have selected one of them and want to change that selection, just click on the selected box to unselect it. Click on the 'Submit' button once you have chosen the two participants.

5	8	3	4
			
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Submit

You will now begin the math test.

Remember:

You have **10 minutes** to solve as many math problems as you can.

Payment: \$1 per correctly solved question, minus \$0.25 per incorrectly solved question.

Two participants in the group (as explained in the previous screens) will get extra \$10 bonus.

When you are ready, please press "Start".

Start

ROUND 2—10 MINUTES. Example:

If $(d-3n)/(7n-d)=1$, which of the following must be true about the relationship between d and n ?

- n is 4 more than d .
- d is 4 more than n .
- n is $7/3$ of d .
- d is 5 times n .
- d is 2 times n .

OK

While we process the bonus winners of the second round, you will be given the chance to earn more money by completing another round of the same task.

Just like in the first two rounds, you will have 10 minutes to solve as many math problems as you can. You will get \$1 for each question you answer correctly, and will lose \$0.25 for each question you get wrong (so nothing can be gained from guessing).

Unlike the second round, this round will not have a bonus payment of \$10 for the top performers.

Once you complete this round, you will get the amount of money you earned in the three rounds, plus the bonus of \$10, if you were one of the winners.

We wait for everyone to be done with round 2, and then we'll give a code to proceed.

Please wait.

You will now begin the math test.

Remember:

You have **10 minutes** to solve as many math problems as you can.

Payment: \$1 per correctly solved question, minus \$0.25 per incorrectly solved question.

When you are ready, please press "Start".

Start

ROUND 3- 10 MINUTES. Example:

If the total surface area of a cube is 24, what is the volume of the cube?

- 8
- 24
- 64
- $48 \cdot (6^{.5})$
- 216

OK

This 10 minutes are over.

Please click on the "OK" button to continue.

OK

Please click OK to begin a brief general questionnaire.
The questionnaire includes questions about the study as well.

OK

What do you think was your score in round 2?
(remember, your score is equal to the number of questions you
answered correctly, minus 1/4 times the number you got wrong)

OK

What do you think was the average score of the other participants in your group in **round 2**?

OK

What do you think are the chances that you won the bonus?

(enter a number between 0 and 1, where 0 means that you are sure you didn't win, and 1 that you are sure you won)

OK

How much effort did you put into the task?

- I didn't put any effort into it 1 2 3 4 5 6 7 I worked as hard as I could

OK

How difficult were the math questions?

- Very easy 1 2 3 4 5 6 7 Very difficult

OK

How do you think the winner selection criterion (i.e., the rule on who gets the bonus in round 2) affected your performance?

It hurt my performance 1 2 3 4 5 6 7 It helped my performance

OK

What do you think were the reasons for the selection policy (i.e., who gets the bonus) in round 2?

Please type your answer in the box below. When you are done click on the OK button to continue.

OK

If you have additional comments, please write them down in the space below.

Please type your answer in the box below. When you are done click on the OK button to continue.

OK

How old are you? ▶

What is your major?

A bat and a ball cost \$1.10 in total. The bat costs \$1 more than the ball. How much does the ball cost?

cents

If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?

minutes

What is your SAT verbal score (a.k.a. critical reading section score)?

What is your SAT math score?

You have completed the experiment!!

Please press the assistance request button to indicate that you are done.

Once everyone is done we'll give you a code to continue to the final screen, where you will find out your total earnings in the study.

Please be patient.

You have completed this experiment.

You made: \$0 in the first round
You made: \$5 in the second round
You made: \$2.75 in the third round
You won the \$10 bonus

Your total earnings are: \$ 17.75

Please write this number down on your payment record sheet, press the assistance request button, and wait for the experimenter.