Bracha, Anat; Cohen, Alma; Conell-Price, Lynn

Working Paper

Affirmative action and stereotype threat

Working Papers, No. 13-14

Provided in Cooperation with:
Federal Reserve Bank of Boston

Suggested Citation: Bracha, Anat; Cohen, Alma; Conell-Price, Lynn (2013) : Affirmative action and stereotype threat, Working Papers, No. 13-14, Federal Reserve Bank of Boston, Boston, MA

This Version is available at:
https://hdl.handle.net/10419/107232

Working Papers

FEDERAL RESERVE BANK OF BOSTON

No. 13-14

Affirmative Action and Stereotype Threat

Anat Bracha, Alma Cohen, and Lynn Conell-Price

Abstract:

In spite of the apparent success of affirmative action (AA) in the past, many oppose such policies. Opponents argue that the cost of attaining proportional representation by preferential policies is too high, reducing the quality of selected groups and stigmatizing members of the protected class. One way in which preferential policies might harm groups they are designed to benefit is by producing *stereotype threat*; that is, cueing a negative stereotype may lead individuals to conform to it. AA, by definition, singles out disadvantaged groups and therefore may unintentionally remind beneficiaries of relevant negative stereotypes and prime a stereotype threat effect. This paper investigates experimentally whether gender-based affirmative action, in the form of a quota, may have this unintended consequence. Using quantitative Graduate Record Examination (GRE) questions, we find that while affirmative action has a positive effect on the performance of non-high-ability women, it negatively affects high-ability women in our sample. Interestingly, there is no evidence that women (of any ability level) in the sample exert less effort under AA or in single-sex competition, or would benefit by doing so. Hence, these findings are consistent with affirmative action having an unintended negative consequence due to stereotype threat. Interestingly, men are not affected by the affirmative action policy regardless of their measured ability in this task. These results should be taken with caution, as this is the first study looking at the stereotype threat effect of AA.

JEL Classifications: C91, J16

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

Anat Bracha is an economist in the research department of the Federal Reserve Bank of Boston. Alma Cohen is an associate professor at the Eitan Berglas School of Economics at Tel Aviv University, the Traphagen Senior Fellow in Law and Economics at Harvard Law School, and a faculty fellow at the National Bureau of Economic Research. Lynn Conell-Price is a graduate student at Carnegie Mellon University. Their e-mail addresses are, respectively, anat.bracha@bos.frb.org, almac@post.tau.ac.il, and lconellp@andrew.cmu.edu.

This paper, which may be revised, is available on the web site of the Federal Reserve Bank of Boston at http://www.bos.frb.org/economic/wp/index.htm.

We would like to thank Uri Gneezy, Muriel Niederle, Analia Schlosser, Tali Regev, and participants of the North American ESA Meetings in Santa Cruz for helpful comments.

This version: September 2013

1. Introduction

The dramatic increase in representation of women in higher education during the 20th century in the United States is strongly associated with affirmative action (AA) programs. During the period when the federal government first promoted affirmative action, female representation increased from about 35 percent of undergraduate students in 1959 to over 50 percent in 1979 (National Center for Education Statistics 2011).

In spite of the apparent success of affirmative action in the past, and the encouragement of such policies today to advance women in the Science, Technology, Engineering, and Math (STEM) fields, where women remain severely underrepresented (NASA 2009, 16–18; NSF 2012), many object to affirmative action policies. The argument against AA policies is that the cost of attaining proportional representation by preferential policies is too high, reducing the quality of selected groups and stigmatizing members of the protected class (for example, S. Steele 1990).

One way in which preferential policies might harm the group they are designed to benefit is by producing a *stereotype threat*. This is an effect whereby reminding someone of a negative stereotype can put the individual at the risk of conforming to the negative stereotype. For example, Shih et al. (1999), using an experimental setting, found that reminding Asian women of their race led them to perform significantly better than with no reminder, while reminding them of their gender led them to perform significantly worse. The stereotype threat effect can be induced by priming—exposing individuals to a stimulus that reminds them of negative stereotypes relevant to them—for example, by asking subjects to indicate their gender (Stricker 1998) or by providing information about differences across identity groups in achievements on the relevant task (Spencer et al. 1999). Affirmative action, by definition, singles out disadvantaged groups and therefore may unintentionally remind beneficiaries of relevant negative stereotypes and prime a stereotype threat effect.

Existing experimental studies such as Schotter and Weigelt (1992), Calsamiglia et al. (2013), Niederle et al. (2013), and Balafoutas and Sutter (2012) do not find a negative effect of implementing affirmative action policies. Yet, the experimental settings considered in these studies are such that stereotype threat is unlikely to occur, potentially neglecting an important

aspect of actual AA policies. In particular, Schotter and Weigelt (1992) use an abstract setting where (cost) disadvantage was randomly assigned, and effort level was captured by choosing a number. Calsamiglia et al. (2013), although studying a real-effort task, use students' performance on Sudoku puzzles across schools where AA is not linked to an established stereotype; rather, AA depended on whether or not one's school has regular Sudoku practice.[1] Finally, Niederle et al. (2013) use math problems to study the effect of preferential gender policies on the decision to participate in a competition; however, the math task was simple, while the gender gap in math performance emerges only with more complex problem-solving (Hyde et al. 1990).

Hence, it is possible that although studies thus far find positive effects of affirmative action, affirmative action in academic admissions, where candidates are judged based on complex stereotype-relevant tasks such as standardized tests, may trigger stereotype threat and result in the unintended consequence of lower performance.

This conjecture is supported by studies showing that stereotype threat affects performance on GRE- and SAT-style questions (Steele and Aronson 1995; Spencer et al. 1999; see Wei 2009 for a summary of lab results) as well as other tasks (Hoff and Pandey 2006; Günther et al. 2010). For instance, Steele and Aronson (1995) demonstrated that, conditional on ability (proxied by SAT scores), black students performed significantly worse than white students on a verbal GRE exam when reminded of their race, while Spencer et al. (1999) find a similar effect on female performance in math. Importantly, gender stereotype threat has been shown to affect performance on actual Advanced Placement Calculus tests (Stricker 1998). Although the majority of studies in the stereotype threat literature find stereotype threat effect, a few recent studies have not (Fryer et al. 2008; Wei 2009, 2012). Wei (2009, 2012) actually finds an opposing effect—stereotype 'reaction'—where females perform best following a stereotype threat prime, and suggests that the *type* of prime used is important to trigger the effect (Wei 2009). This paper adds to the literature by examining whether affirmative action policies serve as a stereotype threat

[1] In addition, Dechenaux et al.'s (2012) survey of experimental research on tournaments mentions an experiment (Michelitch 2009) in which quota-eligible subjects exert less effort in a tournament than quota-ineligible subjects. This work is unavailable in publication, online, or by contacting the author. We therefore do not know the details of the work, including whether it involved real effort.

prime—a type of real-life potential prime that has not been studied before—in an incentivized competitive environment.

To investigate our hypothesis that an affirmative action policy may create a stereotype threat effect we conduct a lab experiment where we implement a gender-based quota policy. Participants in our experiment were assigned into groups of four (two male and two female) and were asked to solve questions taken from the Quantitative Graduate Record Examination (GRE), where males stereotypically (and truly) perform better. Participants were paid for each correct answer and penalized for incorrect answers consistent with the grading of the GRE and other standardized tests used in admission decisions. In addition, participants competed for a monetary prize, to imitate successful school admission, which was awarded either (1) to the top two performers regardless of gender, or (2) 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 or not information on the superior performance (on average) of men compared to women on the actual quantitative GRE exam was presented to participants in AA groups (informational priming[2]). We then analyzed the effect of the gender quota on the performance of men and women by comparing performance (scores on the exam) across the affirmative action conditions, and examined whether AA has a similar effect with and without informational priming.

We find that while affirmative action has no effect on men's performance, it does affect performance for women. Specifically, women of low baseline ability perform significantly better under affirmative action, while women of high baseline ability perform significantly worse. We find no evidence that this pattern is optimal, or that this is consistent with women's response to a single-sex competition. Furthermore, women of all ability levels reported exerting more effort under AA. Taken together, the pattern we find suggests that two opposing factors are at play. The positive effect is driven by the encouragement embodied in higher return to effort (due to its effect on the likelihood of winning the prize), which is positive for women with low-baseline ability and is decreasing with ability. The negative effect on high-ability women is surprising,

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

given that we do *not* find evidence that it is driven by a response to a single-gender competition or by reduced effort. Hence, the results support the hypothesis that AA may have an additional negative effect consistent with stereotype threat.

2. Experimental Design and Procedure

In order to test the effect of affirmative action on tasks that are relevant to admissions in STEM graduate programs, we use a between-subject experimental design with random assignment to gender-based affirmative action in a competitive setting with incentivized performance. The task selected is a test composed of questions from past quantitative GRE exams, since GRE performance is a factor considered in actual admissions decisions. To examine the effect of affirmative action on performance we calculate a score that penalizes guesses—as is standard practice on exams such as the GRE. Specifically, a point was awarded for each correct answer and a quarter point was subtracted for each incorrect answer.

Because we expect that ability may interact with responses 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 incentives, as a proxy for ability. The second-round math exam is the main focus of our analysis. In this round, subjects were randomly assigned to a group of two men and two women[3] and competed with this group for a bonus of $10 on top of their pay for performance. The two group members with the highest scores earned the bonus in the control condition, while in treatment conditions we imposed a gender quota that required at least one woman to be awarded a bonus.

The groups competing in the second-round exam were randomly assigned to one of three conditions: the control condition (No AA), the affirmative action condition (AA), or the affirmative action and informational prime (AA-I) condition. In the control (No AA) condition, the two subjects with the highest scores won the tournament and received the $10 bonus. In the AA and AA-I conditions, the two subjects with the highest scores received the $10 bonus subject to a gender quota that required one woman to receive the bonus. That is, if the two highest scorers were both men, then the highest scorer and the highest female scorer earned the bonus.

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

The AA-I condition is identical to the AA condition except that participants assigned to this condition also received an informational stereotype prime prior to the second-round exam. We include this manipulation in order to compare the effect of the quota policy alone, which may convey information that acts as a stereotype prime, to a direct stereotype threat prime similar to primes used in previous studies (for example, 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 score, as in the first round. After completing the three rounds, subjects filled out a questionnaire before learning their earnings for the three rounds and whether they had won the bonus. In this questionnaire we ask subjects to self-report SAT scores (quantitative and verbal), major, and the extent to which they exerted effort on the exam. After completion of the three rounds and exit questionnaire, participants were informed of their earnings in the three rounds and whether or not they had won the $10 prize, and were paid in private, in cash.

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, with 80 subjects in the control condition, 84 subjects in each of the AA and AA-I conditions, and each condition composed of equal numbers of men and women. All subjects were undergraduate or graduate students from Harvard University recruited from the lab's subject pool. Average earnings were $25.43, and the average age was 20 years old. The (self-reported) average quantitative SAT score was 729.73, the average verbal SAT score was 719.46, and the average major on a one-to-four math-intensive scale was 1.84, where math and engineering is classified as four.

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 test score in the second round. The score is the appropriate measure to examine, as it determines subjects' payment and captures both the quantity and accuracy of their responses. We analyze the effect of AA on the second-round score using a simple ordinary least squares (OLS) and a weighted least squares (WLS) model that adjusts for a systematic relationship between variance in score and the number of questions attempted. To test whether the effect on high-ability women is different from the effect on non-high-ability women, we use test score in the first round as a measure of ability and interact this term with AA. We also explore the effect of AA on the number of questions attempted and response accuracy, measures that may capture changes in response strategy. Finally, we compare the performance of winners of the tournament in round 2 with and without the quota, to check whether the AA policy reduces the quality of the selected group.

3.1 Main Results

Table 1 presents descriptive information on gender differences in performance for each round. In both the baseline round and the second round with the tournament, the average score for men is higher than the average score for women. Specifically, in round 1, men's average score is 6.36, while women's average score is 5.65 (the difference of 0.74 is marginally significant on a one-sided *t*-test with a one sided p-value of 0.08), and in round 2, men's average score is 7.31, while women's average score is 6.45 (the difference of 0.907 has a one-sided p-value of 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 answer significantly more questions in every round, while in two of the three rounds women were slightly more accurate than men. However, the difference in accuracy was not statistically significant.

Given the gender difference in baseline ability and response-strategy, we opted to analyze the effect of AA on second-round scores for each gender separately. In doing so we control for ability using first-round scores and for response strategy using the number of questions

attempted. We also control for age, self-reported SAT scores (verbal and quantitative), and major. Each major was coded on a scale from 1 to 4, representing the quantitative requirements in that field (the higher the number, the higher the quantitative skills required). Table 2 shows how majors are coded.[4]

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 equal to one when the AA policy is in effect and zero otherwise, and a second that also includes an interaction term of the AA dummy variable with baseline ability, as measured by first-round score. Table 3 reports the OLS results of these two specifications. We find that, as expected and regardless of the specification considered, baseline ability (first-round score) has a positive effect on second-round score and is highly significant for both men and women. We also find that the self-reported quantitative SAT score has an additional positive effect on the second-round score. However, the self-reported verbal SAT score and major have nonsystematic explanatory power. The number of questions attempted in the first round, which may capture response strategy,[5] is associated with significantly higher second-round scores for women (effect of 0.455 or 0.479, depending on the specification, both of which have a p-value lower than 0.001). The effect for men is half that for women and is not significant.

Examining the effect of AA on performance, regardless of ability (Table 3, column 2), we find a positive and insignificant effect. However, allowing for the possibility that high- and low-performing women may be affected differently (Table 3, column 4), we find a positive and significant main effect of AA (1.404 with p-value of 0.047), which declines with ability (-0.195 with p-value of 0.054). The rate of diminishing effect of AA is such that the overall effect of AA is negative for high-ability women whose round 1 score is over 7.2. Approximately 75 percent of

[4] While there is no significant gender difference in average quantitative SAT score in our sample (females' average quantitative SAT score is 736, while males' is 718; one-sided *t*-test yield $p_{1\text{-sided}}=0.11$), men's chosen majors are significantly more quantitative-intensive than those chosen by women (the average quantitative score of major chosen by females is 1.87, while for males it is 2.26; this is significant, a one-sided *t*-test yields $p_{1\text{-sided}}<0.001$).

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

women had a score lower than 7.2 in round 1, which implies that AA has a positive effect on most women, but the projected effect of the policy is negative for the top 25 percent.[6] For men, there is a positive effect of AA regardless of ability, but this effect is insignificant.

One technical concern that could affect the results reported above is heteroskedacticity, where the variance in 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. For example, when only one question is attempted, the scores must fall between -0.25 and 1, while when the number of questions attempted is 20 the potential score could be anything between -5 and 20. To test whether the variance is increasing with the number of questions attempted, we conduct a Goldfeldt-Quant test, and the F-test for equality of variance obtained from the test suggests that we can reject this hypothesis in favor of increasing variance (this is true for both regression specifications considered). On the basis of this relationship, we replace the OLS model with a weighted least squares (WLS) model with weights proportional to the number of attempted questions in round 2 (see Table 4).

The WLS results reported in Table 4 are similar to the results reported above, with a slightly stronger effect for women. Examining the effect of AA and allowing for the effect of AA to vary with baseline ability (columns 3 and 4), we find that the main effect of AA on women's performance remains positive (with a coefficient of 1.888 and a p-value of 0.017), but this effect declines with ability. In particular, the interaction of AA with the first-round's score is -0.221 (with a p-value of 0.038), which again implies that the effect of AA on high-ability women (with first round scores of 8.54 and above) is negative. The effect of AA on men is consistent with the results of the OLS regression and remains insignificant.

The gender quota that we implemented objectively increases (at least weakly) a woman's chance to win the bonus. For this reason, the positive main effect of AA on women's scores reported in Tables 3 and 4 may not be surprising. Yet, the fact that the positive effect of AA

[6] 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 the results for the entire sample are.

decreases with ability, to the point that it has a negative effect on high-ability women 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 whether AA alone is enough to obtain this negative effect. We also check whether the negative effect of AA can be explained by high-ability women reducing their effort. We examine both the evidence that high-ability women actually reduce effort, and whether reducing effort could be an optimal response.

Table 5 presents results of WLS regressions, where we add to the specification from Table 4 another variable that accounts for the presence of the informational prime. This variable is a dummy variable equal to one for participants who were informed of the inferior performance of women relative to men on the GRE (on average, 15 percent lower) and zero otherwise. Interestingly, we find no significant effect of the additional information on women, and the results are very similar to the results in Table 4. That is, we find that the effect of AA on women is the same whether or not participants received a direct informational prime, and we find that AA negatively affects the performance of high-ability women. Using the results reported in Table 5 and the distribution of first-round score for women we find that women with first-round scores below 9.09—about 80 percent of the women in our sample—improve their score under the AA treatment, and women with ability in the top 20 percent perform worse under AA.[7] Table 5 shows that the results for men are very similar to those in Table 4, indicating that adding the informational prime dummy did not change the results.

Finding no effect of the informational prime beyond AA is consistent with AA being a prime triggering stereotype threat. However, there is also the possibility that the gender stereotype has no effect at all. It is possible that AA encourages non-high-ability women by sufficiently increasing the marginal benefit for their effort, while at the same time AA makes the

[7] The different specifications lead to a slightly different score threshold for the switch in the effect of AA. Nevertheless, in all specifications the negative effect of AA affects women in the top 20–25 percent.

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 non-high-ability women and negative for high-ability women. Another possibility is that AA leads women to focus on single-sex competition, which may encourage non-high-ability women to compete, consistent with finding that women compete more in a women-only competition (for example, 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 the absolute chance of winning the bonus for women, optimal effort is determined by the marginal effect of effort on the probability of winning the prize. Hence, to determine whether it is optimal for women in our study to increase their effort in response to AA, we must examine how the effect of effort on the probability of winning the bonus changes with AA. That is, if the return to effort in terms of increased winning probability is higher under AA, we would expect higher effort (and therefore score) under AA, while if the return to effort in terms of increased probability of winning is lower under AA, we would expect the policy to result in lower effort. 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 would end up determining whether the particular female considered wins the bonus. Since the single-sex competition is likely more important for women with lower ability, and since the change in their probability of winning without AA is likely lower, it may be that the pattern of performance observed is simply the optimal reaction to AA.

To conclude whether the pattern observed is optimal, we have to determine whether indeed the marginal change in probability due to greater effort is higher for non-high-ability and lower for high-ability women under AA. To do that, we calculate for each woman whether she would win a competition against her other group members, based on first-round performance. We repeat that exercise assuming she solves one more question correctly (as a proxy for exerting

more effort), and 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 winning probability is higher under AA, but the difference is insignificant. Once we split the analysis to examine high- and non-high-ability women separately, we find that for high-ability women the difference is zero, while for the lower-ability women this difference is positive and significant.[8] This calculation suggests that it was not optimal for women of any level of ability in our study to reduce their effort.[9]

Although not optimal, it may be still the case that the women exert less effort in response to AA and that this reduction 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 in response to our questionnaire at the end of the experiment. We run ordered probit regressions of these responses on (1) whether or not the participant was assigned to the AA condition, (2) whether information on the gender gap in GRE performance was provided, (3) first-round score, and the controls of age, self-reported SAT scores (quantitative and verbal), and major (see Table 6). We find evidence that under affirmative action women exert more effort, whether or not they were exposed to the informational prime. These results are robust to adding the interaction of the dummy variable of quota policy with ability (measured by first-round score). For men, we find an insignificant effect of affirmative action and of information; this is true whether or not the effect of AA is allowed to vary with ability.

Hence, even if AA changes women's perception of their probability of winning, they do not decrease their (self-reported) effort in response; in fact, we find evidence that they increase effort across the board.

[8] Since we are using the score in round 1 in our calculation, we determine high-/non-high-ability women by using the reported quantitative SAT scores. We used different thresholds for the high-/non-high-ability classifications—770, 780, 790, and 800. The results are the same no matter which threshold is selected.

[9] Note that the marginal winning probability was calculated by assuming an additional question solved correctly. Examining the average performance change between round 1 and round 2, we find that it is positive and within 1 point. That is, the marginal winning probabilities we calculated are relevant for the effort decision of round 2.

To recap: it is not optimal for women in our study to reduce effort in response to AA, and indeed they do not seem to do that. Hence, the lowered performance of high-ability women under AA does not seem to be explained by high-ability women reducing effort.

3.2.2 Women-Only Competition

To look into the possibility that single-gender competition reduces high-ability women's concern and effort while encouraging the non- high ability women to compete, we ran 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 (it has the same task and bonus for the winner in round 2). 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 were assigned to mixed-gender groups and were not subject to AA. Table 7 shows that there is no significant difference between the women-only condition and the control group. 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-gender group to a women-only group.

Hence, even if AA leads women to focus on competition against another woman, actual scores of women in round 2 do not seem to be negatively affected by it. Taken together, the findings are inconsistent with higher-ability women exerting less effort due to AA, either due to focusing on a competition against one woman or due to higher chances of winning. Hence, the negative effect of AA on high-ability women supports the hypothesis that AA policy evokes stereotype-threat effect.

3.3 What Drives the Change in Performance?

There are two factors that together determine the score. The first factor is the number of questions the participant tried to solve (questions attempted), and the second factor is how well

the participant did (success rate). To test whether AA affected 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 women attempted, we find an insignificant negative main effect of AA (Table 8, column 2), and a negative and statistically significant effect of the informational prime (-0.897 with a p-value of 0.085). For men, we find (Table 8, column 1) the opposite results: a positive and insignificant main effect of AA and a positive and statistically significant effect of the informational prime (1.003 with a p-value of 0.06). 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 number of questions attempted does not fully explain the effect of AA on women that we observe, so we turn to examining the success rate. Table 8, column 4 shows that the main effect of AA is positive and statistically significant (with a coefficient of 0.145 and a p-value of 0.027). We also find that this effect is decreasing with ability (with a coefficient of -0.014 and a p-value of 0.051). In other words, the effect of AA on women is reflected in their success rate in solving the questions.

The pattern we find matches the pattern we observed in the analysis of scores. Namely, while AA does not significantly change the number of questions attempted, it positively affects the success rate for low-baseline-ability women and negatively affects the success rate of high-ability women. Together, the number of correctly solved questions (not shown) significantly increases on average by 1.61 questions (with a p-value of 0.008), and this positive effect declines with ability at the rate of 0.185 (with a p-value of 0.037). Interestingly, for men (Table 8, column 3), we find that the main effect of AA and its interaction with ability on number of questions attempted, success rate, and as a result, number of questions solved correctly, is small and insignificant.

The reduced effectiveness (success rate) rather than reduced effort illustrates nicely the effect of stereotype threat. The fact that we find it among the high-ability women and not among the lower-ability women may be due to the encouragement effect embodied in AA for non-high-ability women. AA gives non-high-ability women a real chance for success that may overcome the negative effect of the stereotype threat. High-ability women, on the other hand, do not enjoy the same encouragement effect, as for them there is no change in the marginal probability of winning with AA, and so they are left with only the negative effect. It is also possible that the non-high-ability women approach quantitative GRE questions with reduced confidence even without the AA prime, yet the high-ability women do not. Hence, the stereotype effect is revealed when looking at the performance of the high-ability women.

3.4. The Effect of Affirmative Action on the Set of Winners

To test whether the quota reduces the quality of winners, which would be analogous to an AA policy leading to admission of a lower-quality group of students, we examine the second-round test scores among the winners, with and without the quota. The average score of winners in the control group was 9.58, while the average score of winners in the AA conditions was 9.29. This difference is not statistically significant (with a p-value of 0.655). Similarly, comparing the average score of actual winners in conditions with AA with the average scores of those who would have won if the quota had not been in effect, we find that the average score among the actual winners was 9.29, while the average score of those who would have won the prize if the quota was not in effect was 9.56; this difference is insignificant (with a p-value of 0.610). Both of these results suggest that the implementation of the quota does not lower the quality of the group of winners.

4. Concluding Remarks

We find that AA has an effect on women's performance on quantitative GRE questions that are used in admission. We find this effect in an incentivized and competitive environment analogous to real-life situations. Interestingly, we find that AA has a positive significant effect on women with lower baseline ability but a negative significant effect on women with higher ability. Examining the marginal return to effort across conditions, we find that reducing effort is

not optimal for women in our sample of any ability level. Indeed, self-reported effort level confirms that AA does not lead to effort reduction. We also find no evidence that single-sex competition between women explains this effect. In other words, high-ability women should not, and do not, seem to reduce effort in light of higher chances to win or the salience of a single-sex competition. Lastly, we find that AA in and of itself has a similar effect whether or not it is accompanied by a direct stereotype prime. These results are therefore consistent with AA being a stereotype prime that may lead to the unintended negative consequence of impairing the performance of the protected group. Nevertheless, since this is the first study that looks at the question of AA and stereotype threat, its implication for the consequences of AA should be taken with caution.

References

Balafoutas, Loukas, 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.

Dechenaux, Emmanuel, Dan Kovenock, and Roman Sheremeta. 2012. A survey of experimental research on contests, all-pay auctions and tournaments. Working Paper

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

Günther, Christina, Neslihan A. Ekinci, Christiane Schwieren, and Martin Strobel. 2010. Women can't jump? An experiment on competitive attitudes and stereotype threat. *Journal of Economic Behavior & Organization* 75(3):395–401.

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

Hoff, Karla, Priyanka Pandey. 2006. Discrimination, social identity, and durable inequalities. *American Economic Review* 96(2):206–211.

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

Michelitch, Kristin 2009. Do quotas make gender and ethnic groups expend less effort in competition? Working Paper.

NASA. 2009. Title IX & STEM: Promising Practices for Science, Technology, Engineering, & Mathematics. Available at http://odeo.hq.nasa.gov/documents/71900_HI-RES.8-4-09.pdf

National Center for Education Statistics. 2011. Historical summary of faculty, enrollment, degrees, and finances in degree-granting institutions: Selected years, 1869–70 through 2009–10 in Digest of Education Statistics.
Available at http://nces.ed.gov/programs/digest/d11/tables/dt11_197.asp

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..

NSF. October 2012. Women, minorities, and persons with disabilities in Science and Engineering. Table 3-1.
Available at http://www.nsf.gov/statistics/wmpd/pdf/tab3-1.pdf

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, S. 1990. *The content of our character: A new vision of race in America*. New York: St. Martin's Press.

Steele, Claude 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.

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.

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 (mean)	Female (mean)	Diff.	Hyp.
Round 1	Score	6.36	5.65	0.74*	>
	# Questions	8.91	8.32	0.64*	>
	Ratio correct	0.76	0.72	0.03	
Round 2	Score	7.31	6.45	0.91*	≠
	# Questions	11.50	9.89	1.65***	≠
	Ratio correct	0.69	0.71	-0.02	
Round 3	Score	7.54	7.06	0.48	
	# Questions	12.20	11.26	0.94**	≠
	Ratio correct	0.68	0.68	<0.001	

Table 2: Coding of Majors, Increasing in Quantitative Requirements

Major type = 1	Major type = 2	Major type = 3	Major type = 4
African American Studies, Anthropology, Art History, Classics, East Asian Studies, Education, English, Government, History, Linguistics, Literature, Music, Near Eastern Lang., Philosophy, Politics, Public Admin., Public Health, Religion, Social Studies, Sociology, and Visual studies.	Architecture, Biology, Earth & Planet Science, History of Science, Neuroscience, and Psychology.	Biochemistry, Chemistry, Computer Science, Economics, and Statistics.	Applied Math, Engineering, Math, and Physics

Table 3: OLS Regressions of 2nd Round Score on AA

	(1) Male	(2) Female	(3) Male	(4) Female
Affirmative Action (AA)	0.414	0.316	0.571	1.404**
	(0.510)	(0.397)	(0.867)	(0.698)
Score in 1st Round (Score R1)	0.609***	0.239***	0.621***	0.367***
	(0.132)	(0.0891)	(0.132)	(0.102)
(AA)x(Score R1)			-0.0242	-0.195*
			(0.126)	(0.0998)
# Questions in 1st Round	0.202	0.455***	0.197	0.479***
	(0.125)	(0.0660)	(0.125)	(0.0722)
Age	-0.141	0.0465	-0.138	0.0124
	(0.0976)	(0.0930)	(0.0971)	(0.0950)
SAT Quantitative	0.0216***	0.0179***	0.0217***	0.0178***
	(0.00509)	(0.00320)	(0.00525)	(0.00311)
SAT Verbal	0.00109	0.000299	0.00107	0.000392
	(0.00372)	(0.00331)	(0.00374)	(0.00331)
Major	-0.133	0.169	-0.132	0.127
	(0.256)	(0.297)	(0.258)	(0.296)
Dummy for Missing SAT-Q	18.27***		18.38***	
	(4.016)		(4.172)	
Dummy for Missing SAT-V	-2.994		-3.060	
	(2.870)		(2.895)	
Dummy for Missing Major	0.281	0.254	0.275	0.159
	(1.095)	(0.749)	(1.098)	(0.736)
Constant	-12.43**	-13.54***	-12.59**	-13.68***
	(5.052)	(3.670)	(5.188)	(3.599)
N	123	123	123	123
r2	0.722	0.748	0.722	0.755

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4: WLS Regressions of 2nd Round Score on AA

	(1) Male	(2) Female	(3) Male	(4) Female
Affirmative Action (AA)	0.467	0.498	0.751	1.888**
	(0.551)	(0.438)	(0.938)	(0.779)
Score in 1st Round (Score R1)	0.634***	0.238**	0.651***	0.387***
	(0.136)	(0.0949)	(0.137)	(0.108)
(AA)x(Score R1)			-0.0393	-0.221**
			(0.129)	(0.105)
# Questions in 1st Round	0.195	0.438***	0.186	0.465***
	(0.133)	(0.0683)	(0.133)	(0.0754)
Age	-0.121	-0.00860	-0.115	-0.0573
	(0.124)	(0.106)	(0.121)	(0.109)
SAT Quantitative	0.0232***	0.0182***	0.0235***	0.0177***
	(0.00554)	(0.00353)	(0.00577)	(0.00340)
SAT Verbal	0.00120	0.00236	0.00117	0.00241
	(0.00406)	(0.00353)	(0.00407)	(0.00345)
Major	-0.109	0.332	-0.103	0.293
	(0.266)	(0.351)	(0.269)	(0.349)
Dummy for Missing SAT-Q	19.54***		19.83***	
	(4.391)		(4.637)	
Dummy for Missing SAT-V	-2.986		-3.104	
	(3.123)		(3.149)	
Dummy for Missing Major	0.799	1.139	0.790	1.046
	(1.150)	(0.879)	(1.154)	(0.849)
Constant	-14.24**	-14.36***	-14.64***	-14.18***
	(5.445)	(4.269)	(5.567)	(4.162)
N	123	123	123	123
r2	0.761	0.770	0.761	0.779

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5: WLS Regressions of 2nd Round Score on AA and Informational Prime Dummies

	(1) Male	(2) Female	(3) Male	(4) Female
Affirmative Action (AA)	0.576	0.614	0.872	2.028***
	(0.650)	(0.510)	(0.996)	(0.752)
Informational Prime	-0.209	-0.248	-0.214	-0.276
	(0.608)	(0.529)	(0.610)	(0.522)
Score in 1st Round (Score R1)	0.629***	0.237**	0.647***	0.388***
	(0.134)	(0.094)	(0.135)	(0.108)
(AA)x(Score R1)			-0.041	-0.223**
			(0.132)	(0.104)
# Questions in 1st Round	0.200	0.439***	0.190	0.467***
	(0.131)	(0.066)	(0.131)	(0.073)
Age	-0.120	-0.013	-0.114	-0.062
	(0.125)	(0.104)	(0.122)	(0.106)
SAT Quantitative	0.023***	0.018***	0.023***	0.018***
	(0.006)	(0.004)	(0.006)	(0.003)
SAT Verbal	0.001	0.002	0.001	0.002
	(0.004)	(0.004)	(0.004)	(0.003)
Major	-0.093	0.334	-0.087	0.295
	(0.268)	(0.354)	(0.270)	(0.352)
Dummy for Missing SAT-Q	19.209***		19.491***	
	(4.613)		(4.851)	
Dummy for Missing SAT-V	-2.822		-2.940	
	(3.139)		(3.167)	
Dummy for Missing Major	0.822	1.086	0.814	0.985
	(1.150)	(0.909)	(1.155)	(0.876)
Constant	-14.148**	-14.169***	-14.554**	-13.966***
	(5.505)	(4.266)	(5.618)	(4.151)
N	123.000	123.000	123.000	123.000
r2	0.761	0.770	0.762	0.780

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Ordered Probit Regressions of Reported Effort on Condition

	(1) Male	(2) Female	(3) Male	(4) Female
Affirmative Action (AA)	-0.070	0.518**	0.098	0.666**
	(0.235)	(0.261)	(0.458)	(0.399)
Informational Prime	0.257	-0.214	0.253	-0.215
	(0.268)	(0.256)	(0.269)	(0.256)
Score in 1st Round (Score R1)	0.122	0.036	0.135**	0.053
	(0.055)	(0.054)	(0.063)	(0.057)
(AA)x(Score R1)			-0.026	-0.026
			(0.063)	(0.052)
# Questions in 1st Round	-0.010	-0.017	-0.104**	-0.014
	(0.049)	(0.050)	(0.049)	(0.051)
Age	0.016	0.014	0.018	0.008
	(0.048)	(0.065)	(0.047)	(0.068)
SAT Quantitative	0.00320	0.005**	0.003	0.005**
	(0.00244)	(0.0024)	(0.002)	(0.002)
SAT Verbal	0.00036	-0.002	0.0003	-0.002
	(0.00176)	(0.002)	(0.002)	(0.002)
Major	0.026	0.006	0.027	-0.0003
	(0.127)	(0.151)	(0.127)	(0.153)
Dummy for Missing SAT-Q	2.243		2.338	
	(2.007)		(2.000)	
Dummy for Missing SAT-V	-0.213		-0.280	
	(1.368)		(1.352)	
Dummy for Missing Major	-0.362	-0.139	-0.367	-0.154
	(0.426)	(0.434)	(0.426)	(0.437)
N	123	123	123	123

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 7: WLS Regressions of 2nd Round Score Comparing Women-Only Condition to AA Conditions

	Control v. Women Only
Women Only	0.306
	(0.909)
Score in 1st Round	0.606***
	(0.126)
(Women Only)x(Score R1)	0.0381
	(0.101)
# Questions in 1st Round	0.251**
	(0.111)
Age	0.0203
	(0.109)
SAT Quantitative	0.00846
	(0.00604)
SAT Verbal	0.00455
	(0.00380)
Major	0.205
	(0.306)
Missing SAT-Q	10.05*
	(5.254)
Missing Major	2.351**
	(0.971)
Constant	-9.857*
	(5.882)
N	73
r2	0.843

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 8: WLS Regressions of 2nd Round # of Questions and Success Rate

	(1)	(2)	(3)	(4)
	# Questions		Success Rate	
	Male	Female	Male	Female
Affirmative Action (AA)	0.324	-0.052	0.025	0.145**
	(0.890)	(0.894)	(0.066)	(0.065)
Informational Prime	1.003*	-0.897*	-0.044	0.026
	(0.535)	(0.517)	(0.037)	(0.032)
Score in Round 1	0.070	0.167	0.032***	0.016*
	(0.137)	(0.149)	(0.008)	(0.008)
(AA)x(Score R1)	-0.032	-0.033	0.002	-0.014*
	(0.106)	(0.115)	(0.007)	(0.007)
#Questions in 1st Round	0.629***	0.493***	-0.017**	0.004
	(0.143)	(0.163)	(0.008)	(0.007)
Age	-0.099	0.099	-0.004	-0.003
	(0.119)	(0.194)	(0.008)	(0.008)
SAT Quantitative	0.012**	0.010**	0.002***	0.001***
	(0.005)	(0.004)	(0.000)	(0.000)
SAT Verbal	0.002	-0.005	0.000	0.000
	(0.004)	(0.004)	(0.000)	(0.000)
Major	-0.079	0.190	-0.006	0.001
	(0.240)	(0.318)	(0.017)	(0.021)
Missing SAT-Q	12.589***		1.259***	
	(4.402)		(0.337)	
Missing SAT-V	-2.946		-0.176	
	(2.810)		(0.213)	
Missing Major	-0.056	2.443**	0.030	-0.038
	(0.844)	(1.206)	(0.068)	(0.056)
Constant	-2.505	-0.880	-0.465	-0.859**
	(4.887)	(5.366)	(0.400)	(0.354)
N	123.000	123.000	123.000	123.000
r2	0.685	0.648	0.505	0.454

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendix:

If the model is:

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

$$max_e\, p(e;AA)Bonus + \ Reward(e) - c(e),$$

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

The probability of winning the bonus when AA=0 is:

$$p_1(e) + [1 - p_1(e)]p_2(e).$$

The probability of winning the bonus when AA=1 is:

$$p_1(e) + [1 - p_1(e)]p_2(e) + [1 - p_1(e)][1 - p_2(e)]qw_1(e).$$

The change in probability as effort (e) increases:

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

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

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

$$p'(e; AA = 0) \lessgtr p'(e; AA = 1) \leftrightarrow$$

$$p_1'(e)[1 - p_2(e)] + [1 - p_1(e)]p_2'(e) \lessgtr$$

$$p_1'(e)[1 - p_2(e)][1 - qw_1(e)] + [1 - p_1(e)]p_2'(e)[1 - qw_1(e)]$$

$$+[1 - p_1(e)][1 - p_2(e)]qw_1'(e).$$

That is:

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

This can be seen as:

(1) $\underbrace{p_1'(e)[1 - p_2(e)] + [1 - p_1(e)]p_2'(e)}_{p'(e; AA = 0)} \lessgtr \underbrace{[1 - p_1(e)][1 - p_2(e)]\frac{w_1'(e)}{w_1(e)}}$

Increase in the chance to win against the other woman weighted by the relative chance of not being in the top two to the chance of being the top woman

OR:

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

For a given effort level and for high-ability women, their chance of not being among the top two is lower than the chance of the non-high-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 non-high-ability women the opposite holds.