The incentive effects of affirmative action in a real-effort tournament

Caterina Calsamiglia a,b, Jörg Franke c,1, Pedro Rey-Biel a,b,⋆

a Universitat Autònoma de Barcelona, Department d'Economia i d'Historia Econòmica, 08193, Bellaterra, Barcelona, Spain
b Barcelona GSE, Spain
c Department of Economics and Social Science, TU Dortmund, 44227 Dortmund, Germany

A R T I C L E   I N F O

Article history:
Received 2 June 2011
Received in revised form 13 November 2012
Accepted 13 November 2012
Available online 23 November 2012

JEL classification:
C72
C91
J78
M52

Keywords:
Affirmative action
Tournament
Real-effort
Experiment
Sudoku

A B S T R A C T

Affirmative action policies bias tournament rules in order to provide equal opportunities to a group of competitors who have a disadvantage they cannot be held responsible for. Its implementation affects the underlying incentive structure which might induce lower performance by participants, and additionally result in a selected pool of tournament winners that is less efficient. In this paper, we study the empirical validity of such concerns in a case where the disadvantage affects capacities to compete. We conducted real-effort tournaments between pairs of children from two similar schools who systematically differed in how much training they received ex-ante on the task at hand. Contrary to the expressed concerns, our results show that the implementation of affirmative action did not result in a significant performance loss for either advantaged or disadvantaged subjects; instead it rather enhanced the performance for a large group of participants. Moreover, affirmative action resulted in a more equitable tournament winner pool where half of the selected tournament winners came from the originally disadvantaged group. Hence, the negative selection effects due to the biased tournament rules were (at least partially) offset by performance enhancing incentive effects.

Crown Copyright © 2012 Published by Elsevier B.V. All rights reserved.

1. Introduction

In many selection processes such as university admissions, job promotions and procurement auctions, competition helps identifying the highest-ability candidates, facilitates the efficient allocation of talent, and provides incentives for individuals to improve their performance. These objectives may not be achieved if some otherwise competitive candidates do not stand a fair chance to win the competition. For example, talented students from poor economic backgrounds may have attended schools that receive less funding which, irrespective of their talent, may affect their SAT performance and hence their university admission. Likewise, some individuals may belong to groups which historically have suffered discrimination, and have to overcome major obstacles in order to be on an equal footing to compete.

Affirmative Action (AA) policies take proactive steps to provide equal opportunities to discriminated groups that have a potential disadvantage. They are often implemented by biasing tournament rules in order to increase the probability of success of an otherwise disadvantaged group. For example, a fixed lump-sum bonus of 20 (out of 150) points was added to the score of minority applicants to the undergraduate admission. Likewise, some individuals may belong to groups which historically have suffered discrimination, and have to overcome major obstacles in order to be on an equal footing to compete.
program at the University of Michigan and a similar but “unofficial lift” scheme is used at several top universities.7 On the other hand in public procurement auctions, bid preferences are granted in a multiplicative way. For example, road construction contracts in California are auctioned off by granting a 5% reduction of the submitted bid to small business enterprises.

In this paper we provide experimental evidence on how AA policies can affect individuals’ effort through a series of real-effort tournaments where pairs of children from two different schools compete against each other in solving simple numerical puzzles known as “sudokus.”4 A key element of our experimental design is that one competitor in each pair faces a naturally induced disadvantage in the respective task.5 This difference is due to the fact that students in one school (“experienced”) are taught how to solve sudokus as part of their regular math classes, while students in the other school (“non-experienced”) are not. The schools are very similar in all other relevant aspects: both are private, located in similar neighborhoods, fully bilingual and have good records in national math and science competitions. Therefore, the difference in experience can be regarded as an exogenous source of disadvantage imposed on one of two similar groups, since solving sudokus in the experienced school was not yet common at the time most parents made their schooling decision.6 We implement two different types of AA policies, lump sum and proportional bonuses, to compensate for this disadvantage in capacities to compete.7 Hence, we analyze affirmative action policies in a situation where a real and systematic disadvantage between competing individuals is present.8

From a theoretical perspective affirmative action policies can be interpreted as discretionary handicapping of specific groups of agents. The incentive effects of AA rules have been theoretically analyzed for rank-order tournaments in Lazear and Rosen (1981), as bid caps in all-pay auctions in Che and Gale (1998) and for revenue-maximizing auctions in Myerson (1981). The explicit analysis of incentive effects of affirmative action policies in various competitive situations has been carried out in Balart (2009), Franke (2012), Fu (2006), and Hickman (2010).9

Our tournament is in line with Schotter and Weigelt (1992), who studied the incentive effects of AA in a laboratory experiment where the choice of a monetary cost symbolizes the choice of exerted effort in a tournament. In their experiment, subjects’ exogenous disadvantage was induced by assigning different cost parameters for which individuals were later compensated for through affirmative action, implemented as lump-sum bonuses of different sizes. Their results indicate that AA can either boost or worsen performance depending on the relative size of the cost disadvantage with respect to the implemented compensation. However, in many situations in which affirmative action is implemented the comparison of sizes in cost disadvantages and compensations is not as straightforward as in a laboratory study guided by an abstract theoretic model. Hence, our experiment complements their study by showing that AA can enhance effort exertion and therefore performance of participants in a real effort framework where the disadvantage is due to a de-facto asymmetry in capacities to compete.10

Subjects in our experiment were school children and the tournament was conducted in their respective schools. Working with children in their natural environment offers a clean setup to analyze students’ behavior when exposed to exogenously different experiences imposed by their schools. Using children as subjects has additional advantages: they react rationally, spontaneously and in line with economic theory; their performance is not affected by them questioning the underlying motivation of the experimentalist; and it is relatively easy to provide them with incentives, see Harbaugh et al. (2001) and Harbaugh and Krause (2000). Finally, studying how children react to AA is important since some social asymmetries may be ideally resolved at early ages, before they are exacerbated.

Based on two benchmark treatments we first confirm the crucial assumption of our set up, that is, that the asymmetry in experience is reflected in subjects’ performance. In fact, subjects from the school where solving sudokus was part of regular class activities had an advantage in the competition: they solved significantly more sudokus than non-experienced subjects in a benchmark treatment where this difference in experience was not revealed to the tournament participants. Moreover, making this asymmetry in experience publicly known did not lead to significant changes in performance of neither experienced nor non-experienced individuals.

Regarding the incentive effects of AA we find some evidence that performance of non-experienced subjects was enhanced. However, for experienced individuals performance effects differed with ability: affirmative action enhanced the performance of most subjects where this effect declined in ability such that subjects of highest ability actually lowered their performance. Those effects were relatively robust with respect to the two different types of AA policies we implemented in

3 This procedure was recently ruled to be unconstitutional by the Supreme Court, due in part to alleged discriminatory effects on incentives that such compensation may create. State-funded universities such as California, Florida and Texas have also applied similar policies in the past. A number of papers have analyzed the effects of such banning on the efficiency of care-blind policies that result from colleges preferences for diversity. See, among others, Chan and Oyster (2003) and Fryer and Louy (2010).

4 Some recent experimental studies analyze the consequences of different types of naturally occurring differences among social groups in competitive situations. For instance, Holf and Pandey (2006) study stereotype threats among members of different social castes in India. Geerz and Vestrelund (2007) study the different propensity to compete of women and men in mixed-gender tournaments. A recent study that analyzes AA based on the mentioned gender difference is Niederle et al. (forthcoming), where their focus is on the effects of gender quotas on women’s participation decisions and performance in tournaments.

5 Coate and Louy (1993) show theoretically how discrimination may arise in two symmetric groups as a self-fulfilling prophecy when a group expects to be discriminated. Our experimental design includes a treatment in which subjects are aware that a difference in experience exists (K) and another in which they are not (NK). We do not observe differences in performance between both treatments, showing that non-experienced subjects, who could feel discriminated against when knowing they are less experienced do not react as predicted by their model.

6 We do not compare behavior between schools but between treatments in the same school. Hence, similarity between schools is not a necessary condition for the validity of our analysis. However, the justification of affirmative action is based on the premise that the only relevant difference between schools is the ex-ante experience in solving sudokus.

7 Calsamiglia (2009) shows that an appropriately designed AA policy should equalize rewards to effort whenever the set-up affects one of many factors determining individual welfare. In this particular environment rewards to effort are equalized with proportional AA.

8 An important strand of the literature has investigated the fairness and efficiency effects of implementing AA policies. Holzer and Neumark (2000) survey this literature and argue that, while AA increases representation, results with respect to the performance of the selected group are mixed. More recently, McCrary (2007) shows that AA increases the representation of blacks in the police force, but that their performance does not decrease in a significant way. Bertrand et al. (2012) analyze the effects of AA in college both in terms of the increased representation and their success in college and in the labor market. Miller and Segal (2012) study the effect of AA on the pool of hired law enforcement officers.

9 One common assumption in these models is that agents are ex-ante heterogeneous in ability or valuations (e.g., due to past discrimination). In Scotcharmer (2008) this assumption is relaxed in the sense that agents are (in expected terms) homogeneous in abilities but groups systematically differ in the variance of their ability distribution. In a promotion context this results in overrepresentation of the group with the noisier signal although this group will have lower average ability. Moreover, it is shown that efficiency and lack of affirmative action are incompatible in this context. In her model though, individuals do not make any effort choice and increased efficiency under AA can be attributed to the fact that it tends to neutralize the bias from increased noise and allows the planner to attract subjects with underlying higher ability.

10 The implementation of AA might also alter the participation decisions of competitors which additionally affects effort incentives for participating agents. Niederle, Segal, and Vestrelund (forthcoming) and Balafoutas and Sutter (2012) find that affirmative action increases the participation of women into gender competitions. Marion (2007) and Krasnokutskaya and Seim (2010) analyze how bid preference programs for minority-run businesses in public procurement auctions can affect the set of participants and the final cost paid for a given project. Our experimental design is based on compulsory pair-wise tournaments which implies that we can abstract from confounding effects due to simultaneous changes in subjects’ participation decisions and therefore can concentrate our analysis on the respective incentive effects.
our tournament, i.e., lump-sum and proportional compensation, but were more pronounced for those policies where the bias was relatively high. We also find that participant’s expectations on the probability of winning were affected by affirmative action consistently with their observed performance.

Moreover, AA policies balanced the pool of tournament winners on average, since around half of the non-experienced subjects in the AA treatments won their respective tournaments in all potential matches. Also, the average performance of all potential tournament winners selected through AA, which naturally includes a higher proportion of less experienced subjects was only marginally lower than the average performance of the winners who would have been selected without it. Hence, the negative effect of selecting a higher proportion of non-experienced individuals as tournament winners was partially compensated by increased levels of effort performed under AA.

The rest of the paper is organized as follows. Experimental design and procedures are explained in Section 2. Section 3 presents the results. Section 4 sums up our conclusions. The Appendix A contains an English translation of the instructions used in the experiment.

2. Experimental design and procedures

We conducted pair-wise tournaments among 336 school children, aged 10–13, from two similar non-religious, bilingual private schools located in similar neighborhoods in Barcelona. Students at both schools had a systematic difference in experience in a specific real-effort task consisting in solving simple “sudokus”. This ex-ante difference in experience was due to the fact that during regular math classes students in the “experienced” school (E) were trained in solving sudokus (and in fact had to solve sudokus as part of their regular homework) while students at the “non-experienced” school (NE) were not.11

Sudoku is a logic-based number-placement puzzle. The objective is to fill a $9 \times 9$ grid so that each column, each row and each of nine $3 \times 3$ boxes contains one-digit numbers from 1 to 9 only once. The puzzle setter provides a partially completed grid. We use a simplified $4 \times 4$ grid version in order to obtain sufficient variation in performance. We chose this task because the rules are simple, yet it requires substantive logical reasoning and concentration by the subjects. Additionally, performance is easy to measure and, crucially, depends on effort. Most importantly, both effort and ability play a role, so that non-experienced subjects still have a chance of winning, independent of whether they are favored by an affirmative action policy or not.12

Fig. 1 below shows one of the sudokus used in the experiment (a) and its solution (b).

Each student from E was randomly and anonymously matched with a student from NE in his or her same school year (4th or 6th grade). Each pair competed in a tournament which lasted 30 min.13 Subjects had to correctly solve as many sudokus as possible in order to beat their matched rival. All subjects were handed the same answer sheet containing 96 sudokus randomly generated with the same level of difficulty by a computer program.14 Each pair of subjects was competing for a 7€ voucher from a bookshop located in Barcelona.15 In each pair, the student who had correctly solved more sudokus than the respective opponent during a 30 min period won the voucher. In the case of ties, the winner was determined randomly.

Once we verified that experience provided an advantage for subjects in the experienced school, our objective was to study: 1) the effect of providing information on competitors’ previous experience with the task; and 2) the effect of implementing affirmative action policies on subjects’ performance and as a result, on the output generated by subjects selected as tournament winners. Thus, we randomly assigned similar numbers of subjects from each school to each of six treatments. In treatment NK no subject was informed about whether subjects from the other school were experienced or not in solving sudokus. In treatment K students at the NE school were told that students in the E school had previous experience in solving sudokus. Similarly, students in the E school were told that students at the NE school were not taught how to solve sudokus. In the remaining four treatments all subjects were informed about the existence of a difference in experience across schools and about the particular affirmative action policy applied to compensate NE subjects. In treatments LH (Lump-sum High) and LL (Lump-sum Low), all subjects were informed that NE subjects were given a predetermined number of solved sudokus ex-ante: 20 in LH

11 An ex-post experimental questionnaire showed that a small fraction of students from both schools were familiar with sudokus due to prior experience outside school. We control for this ex-ante experience in our analysis by using a proxy for ability. The task was defined as “filling in a grid” and the word “sudoku” was never mentioned.

12 In fact, the percentage of NE winners in their respective tournament was at least 13.3% (for experimental treatment “K” and 4th year students, where no affirmative action was implemented).

13 We chose pair-wise tournaments instead of multiple-prize tournaments with N players because the schools did not allow us to establish intra-school competition. Additionally, pair-wise tournaments allowed us to control the amount of information that each subject had on its rival’s ability.

14 The software used was “SuDoku Pro” by Dualogy Systems. The proportion of mistakes across all solved sudokus was not statistically significant. No subject was able to complete all 96 provided sudokus.

15 Subjects were explicitly told that the voucher was redeemable for “books, collector’s cards, toys, music or comics”. Experiments took place at approximately the time the final Harry Potter book, priced at 20 Euro, was published in Spain.
We later assigned subjects to treatments in such a way that the groups were balanced regarding subjects’ gender, birth date, teaching group and school grades. We then randomly matched participants from both schools in order to guarantee their collaboration and pedagogical effectiveness of the treatments. The lump-sum treatments for an average non-experienced participant, and PH bonuses were designed to be close to the respective values for the same treatments at each school.16 Small variations across treatments were made it clear that each student was competing against an anonymous student from another comparable school and that students at the other school were systematically experienced (or not) in solving sudokus (for treatment NK this information was omitted). The difference in ex-ante experience was explicitly mentioned to justify the implementation of the affirmative action bias in favor of the non-experienced group in the AA treatments (see the Appendix A for the instructions). Tournament rules were explained giving numerical examples (specific to each treatment) for all potential outcomes of the tournament, i.e., losing, winning, and tying. Moreover, aggregate information with respect to the number of sudokus (i.e., mean, minimum and maximum) that had been correctly solved by a comparable subject pool was provided.17 This information, identical for all subjects, was based on the results of our pilot experiments. The experimentalist also held up a 7€ voucher to increase the credibility of the prize offered to tournament winners. After that, subjects had 30 min to solve the sudokus in two separate handouts. After the first 15 min, subjects were instructed to start working on the second handout, so that we could measure whether there were intra-session learning effects or whether these were over-ruled by fatigue. Subjects were explicitly told that they could stop solving sudokus at any time and start any other activity, such as drawing, under the condition that they remained quiet and did not bother others. In fact, very few subjects were observed to choose alternative activities. After the 30 min had passed, the handouts were collected and a questionnaire about previous experience in solving sudokus, self-confidence and the perceived fairness of the implemented affirmative action policies was distributed. Once the questionnaires had been filled in, subjects continued with their regular classes. The experimentalists then randomly matched participants from both schools, determined the winners and deposited the vouchers at the schools, so that they could be distributed by school faculty.

Before describing the main results of our experiment, we first justify two important aspects of our design: our measure of individual ability and possible differences in tournament performance between the two fifteen minute periods.

16 We are unaware of cross-contamination between schools or between subjects from different school years at the same school. The timing of the experiments was carefully designed so as to avoid these problems.
17 This was the reason why different treatments were carried out at different time-intervals. Since the experiment deals with effort motivation and children may be easily influenced, it was crucial to have the same experimentalist conducting the sessions. The experimentalist rehearsed repeating exactly the same cues across sessions.
18 We will use the number of correctly solved sudokus in this practice round ("Pretest") as a proxy variable for individual ability. In particular, we use it to control for relative ability levels between subjects among the same school in a treatment. There is a strong positive correlation between this variable and tournament performance (always above 0.67 in our treatments).
21 The empirical analysis revealed that the specific numbers mentioned in this information did not lead to an anchoring effect for the subjects. An anchoring effect would imply, that subjects’ performance in the benchmark treatments would be clustered closely above 25, the mean, while only few observations should be located directly below 25.

Table 1

<table>
<thead>
<tr>
<th>Description of treatments.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Known NK Subjects unaware of others’ experience</td>
</tr>
<tr>
<td>Know K Subjects aware of others’ experience</td>
</tr>
<tr>
<td>Lump-sum High LH Subjects aware of experience and NE subjects receive a bonus of 20 correct sudokus bonus</td>
</tr>
<tr>
<td>Lump-sum low LL Subjects aware of experience and NE subjects receive a bonus of 8 correct sudokus bonus</td>
</tr>
<tr>
<td>Proportional High PH Subjects aware of experience and NE subjects receive 1 correct sudokus bonus for every 1 correct</td>
</tr>
<tr>
<td>Proportional Low PL Subjects aware of experience and NE subjects receive 1 correct sudokus bonus for every 2 correct</td>
</tr>
</tbody>
</table>

and 8 in LL. In treatments PH (Proportional High) and PL (Proportional Low), all subjects knew that NE subjects were given a number of solved sudokus proportional to the number of sudokus they correctly solved, one for every correctly solved sudoku in the case of PH, and one for every two correctly solved sudokus in the case of PL. Table 1 summarizes our treatment design.

Based on this treatment design our analysis proceeds as follows. The comparison of performance in treatment NK between schools E and NE allows us to see whether students from the E school were in fact advantaged due to their ex-ante experience. Comparisons of performance across treatments NK and K within a school allow us to study the effects of making the information on different levels of ex-ante experience salient. Finally, through the comparison of within-school performance between treatments K and the different treatments in which AA is implemented (LL, LH, PL, PH) we can assess the isolated effect of affirmative action, abstracting from the information effect. In parts of our analysis, we pool the data from all treatments where affirmative action is implemented and refer to them generically as the “AA” treatment.

The sizes of the affirmative action policies were determined using results from pilot experiments with children of similar age in other schools where solving sudokus was never part of regular course activities. In such pilots we adopted a 15 min time period without any AA policies and asked for private ex-ante experience with sudokus. The difference in the number of correctly solved sudokus between privately experienced and non-experienced subjects was roughly 4 sudokus. In the main experiment we doubled the time span to 30 min and set the LL bonus such that this difference in experience would have roughly been leveled out. The LH bonus of 20 sudokus was close to the average number of correctly solved sudokus by our pilot non-experienced subjects. The PL and PH bonuses were designed to be close to the respective values for the lump-sum treatments for an average non-experienced participant.

Prior to conducting the experiments, we repeatedly met with faculty from both schools in order to guarantee their collaboration and pedagogical interest in the project. During these meetings we obtained information on subjects’ gender, birth date, teaching group and school grades. We later assigned subjects to treatments in such a way that the groups were balanced in accordance with these pre-specified characteristics. Table 2 below shows descriptive statistics of subjects assigned to each treatment at each school.16 Small variations across treatments were mainly due to absent students and latecomers.17 Participation was mandatory, which helped to avoid selection biases and simplified matters for the school. None of the subjects manifested opposition to participating.

Experiments were carried out on two separate but close dates in 2008. In each school experimental sessions took place at different times of the day for 4th and 6th graders for practical reasons. Subjects were conducted to separate classrooms according to our predefined assignment. While students waited for the experimentalist, teachers conducted a specific and identical school activity (writing an essay) in order to keep the subjects calmed and equally uninformed about the experiment. The same experimentalist arrived at each of the classrooms at twenty-minute intervals and then sessions started. Teachers were not present during the experimental sessions, in order to minimize their influence.

The experimental sessions lasted 1 h. First, the experimentalist read out general instructions on how to solve sudokus (see “Pre-instructions” in the Appendix A). Then, subjects had a five-minute practice round to solve sudokus with no incentives being offered and no mention of competition. After this period, the experimentalist solved one of the practice sudokus in front of the students. Once questions were clarified, instructions for each of the treatments were read aloud. The instructions made it clear that each student was competing against an anonymous student from another comparable school and that students at the other school were systematically experienced (or not) in solving sudokus (for treatment NK this information was omitted). The difference in ex-ante experience was explicitly mentioned to justify the implementation of the affirmative action bias in favor of the non-experienced group in the AA treatments (see the Appendix A for the instructions). Tournament rules were explained giving numerical examples (specific to each treatment) for all potential outcomes of the tournament, i.e., losing, winning, and tying. Moreover, aggregate information with respect to the number of sudokus (i.e., mean, minimum and maximum) that had been correctly solved by a comparable subject pool was provided. This information, identical for all subjects, was based on the results of our pilot experiments. The experimentalist also held up a 7€ voucher to increase the credibility of the prize offered to tournament winners. After that, subjects had 30 min to solve the sudokus in two separate handouts. After the first 15 min, subjects were instructed to start working on the second handout, so that we could measure whether there were intra-session learning effects or whether these were over-ruled by fatigue. Subjects were explicitly told that they could stop solving sudokus at any time and start any other activity, such as drawing, under the condition that they remained quiet and did not bother others. In fact, very few subjects were observed to choose alternative activities.

After the 30 min had passed, the handouts were collected and a questionnaire about previous experience in solving sudokus, self-confidence and the perceived fairness of the implemented affirmative action policies was distributed. Once the questionnaires had been filled in, subjects continued with their regular classes. The experimentalists then randomly matched participants from both schools, determined the winners and deposited the vouchers at the schools, so that they could be distributed by school faculty.

Before describing the main results of our experiment, we first justify two important aspects of our design: our measure of individual ability and possible differences in tournament performance between the two fifteen minute periods.

16 Average Grade is calculated using grades in all topics in the preceding term. It is slightly higher at the NE school than at the E school (3.55 vs. 3.34, significant at the 1% level). This difference may be due to different grading systems across schools.
17 If there were unequal number of participants in one treatment in the two schools we randomly paired unmatched subjects with matched subjects to determine whether they won the tournament. This new match was only used to determine whether unmatched subjects won the tournament or not. Therefore each subject could only win once. This is also the reason why the instructions only specify the consequences for the respective subject and not for his/her direct rival.
2.1. Performance in the non-incentivized trials as a measure of ability

Our main control of individual ability is subjects' number of correct answers in the five minute trial round conducted prior to the experiment, which we will denote as “Pretest” (7 levels for experienced subjects, 13 levels for non-experienced). Since this trial round was not incentivized, a possible concern is that participants, especially those with prior training in the task, would not exert much effort, and thus the interpretation of the trial performance as a measure of ability would be misleading. Fig. 2 below shows the average number of correctly solved sudokus in the tournament by subjects, classified according to their ability levels as measured by their performance in the trial period. From Fig. 2 (A) it is clear that subjects with higher performance in the trial period also performed better in the actual experiment. Moreover, similar observations can be made for each treatment separated in Fig. 2 (B), which includes the two treatments without affirmative action in the tournament (K and NK). This observation is confirmed by calculating the correlation coefficients between trial performance and tournament performance in the respective treatments. All coefficients are high (min. 0.68) and similar across treatments (max. 0.84), indicating that those subjects who solved more correct sudokus in the pretest are also the best performers in the experiment.

2.2. First and second round behavior

The exertion of effort and therefore the respective performance of a subject in the tournament might be time dependent and vary over the duration of the tournament. This can be caused, for example, by either fatigue or learning during the experiment. If this would be the case then the implementation of affirmative action might potentially have different effects during the course of the tournament that are hard to isolate. However, subjects' performance during the tournament did not seem to change substantially. This is suggested by the graphs in Fig. 3, where the average number of correct answers during the first and second fifteen minute periods of the tournament are presented, classified by ability level in Fig. 3(A), and additionally separated by treatment in Fig. 3(B). These graphs give a first indication that treatments did not affect subjects' performance differently in the first and second part of the tournament. Experienced subjects completed on average one more sudoku in the second part than in the first part (significant at the 1% level). Non-experienced subjects did not solve a significantly different number of sudokus in the two parts. Separating by treatment, we do not observe significant differences between both parts of the experiment in any of the treatments. Thus, results indicate that the effects of learning and fatigue may not be relevant or possibly cancel out.

23 There is only one non-experienced subject who correctly solved 11 sudokus in the trial period whose tournament performance clearly differs from the rest of participants.

We can summarize our discussion so far in the following way: firstly, results from the non-incentivized pretest trial are a feasible proxy for individual abilities; secondly, there is no evidence for systematic differences in performance between the first and the second half of the tournament. Hence, in the next section we will use the number of correctly solved sudokus during the complete tournament as a measure for overall performance and control for unobserved individual ability by using the respective individual pretest results.

3. Results

3.1. Descriptive statistics

We start by taking a descriptive look at the data. Table 3 reports the average number of correct sudokus by treatment and school year (4th

![Fig. 2](image-url)
or 6th grade) in each of the schools (E and NE), as well as standard deviations (in parentheses). There is high heterogeneity in performance in all treatments and schools and thus, standard deviations are relatively large. Table 3 provides a first indication that experienced subjects (E) solve, on average, more sudokus than non-experienced subjects (NE), a key hypothesis justifying our experimental design. Aggregating over all treatments, experienced subjects solved 36.06 sudokus, while non-experienced subjects solved 23.31 sudokus.\(^{24}\) Using the number of sudokus solved in the five minute practice round as a measure of individual ability in solving sudokus, we find that experienced subjects of low ability, those who solve three sudokus in the practice round, solve a similar number of sudokus (25.61) in the tournament as the average of non-experienced subjects.

Table 3 also shows that age affected performance. The average performance of 4th grade experienced subjects in all treatments is similar to that of 6th grade non-experienced subjects. Fig. 4 below shows the cumulative distribution functions (CDFs) of tournament performance, i.e., the number of correctly solved sudokus in the E and NE school in the two benchmark treatments NK and K. Note that the distributions have a large spread and range from 0 correctly solved sudokus to more than 70. Stochastic dominance of the CDFs for the E school clearly shows that the lack of experience in solving sudokus is in fact a disadvantage for the NE subjects. Mann–Whitney tests comparing the inter-school number of correct sudokus in both of these treatments show significant differences at the 1% level (p-values of 0.002 for NK and of 0.004 for K).

Intra-school comparisons across treatments are less clear-cut. Table 3 reports the average number of correctly solved sudokus for each of the four separate AA treatments (LH, LL, PH and PL), as well as for the pooled treatment (AA).\(^{25}\) Although standard deviations are very large, the reported averages give a first indication that performance may be

\(^{24}\) Moreover, non-experienced subjects solved on average less sudokus than experienced subjects in each treatment separated, including treatment NK, which is a clear indication that the disadvantage was real and not simply a self-fulfilling expectation.

\(^{25}\) Kruskall–Wallis tests comparing the four AA treatments in each school do not show significant differences among treatments. Thus, we will use the pooled AA treatment in this section.
enhanced when providing information on the difference in experience (K vs. NK treatments) and that affirmative action policies also enhanced performance (AA vs. K treatments, with the only negative comparisons being for E subjects in 4th grade and NE subjects in 6th grade). Fig. 5 depicts the CDFs tournament performance in the two benchmark treatments NK and K, as well as the pooled AA treatment separately for E and NE subjects. Visually, the CDF for the K treatment “almost stochastically dominates” the CDF for the NK treatment in both graphs, suggesting that the provision of information about the disadvantage for one group of subjects did not decrease performance. Similarly, the CDF for the AA treatment also lies mostly below the CDF for the K treatment in both schools which implies that subjects faced with AA policies do not decrease their performance. Comparing the distributions of all treatments based on pair-wise Mann–Whitney and Kolmogorov–Smirnov tests does not generally result in significant differences at the standard levels. This can be attributed to the small sample size in combination with large heterogeneity among subjects, which cannot be controlled for by using non-parametric methods. Therefore, in the next section we rely on regression analysis because it allows us to incorporate substantial additional information on the subject level such as gender, age and two controls of individual ability (performance in the trial round and average grades obtained in school) to control for individual heterogeneity.

### 3.3. The effects of affirmative action on performance

We now study the incentive effects resulting from the implementation of AA policies in separate regressions for the two subject pools of experienced and non-experienced subjects. Table 5 presents the results from OLS regressions with robust and clustered (by treatment and class) standard errors for experienced subjects while Table 6 contains the results for non-experienced subjects. Our baseline treatment in both cases is the benchmark treatment K in which subjects are aware of the asymmetry in experience. Explanatory variables are the four AA treatment dummies, which are additionally pooled in separate regressions, in addition to the other controls already used in regressions presented in Table 4 such as the proxy for unobserved ability (“Pretest”), school grades (“Grade”), and a dummy for being in sixth grade instead of fourth grade (“Year”). As can be seen in Tables 5 and 6, these variables have the expected positive and significant coefficients in almost all specifications. Since subjects might react differently to implemented AA policies depending on their respective ability we include an additional interaction term (“AA*Pretest”) to capture these effects in a linear form (REG (7) and (8) for experienced subjects and REG (11) and (12) for non-experienced subjects). The

---

**Fig. 3.** (A): Tournament performance by ability level in the first and second part of the experiment for the group of experienced subjects (top) and non-experienced subjects (bottom). (B): Tournament performance by ability level separated by treatment in the first and second part for the group of experienced subjects (top) and non-experienced subjects (bottom).

---

26 The range and the variance of the dependent variable are sufficiently large such that OLS regressions should be appropriate. Moreover, our specification is robust with respect to the use of Poisson regressions: results are similar with respect to sign and significance levels of the coefficients (the size of coefficients is not directly comparable between these two approaches as marginal effects in Poisson models are not constant).

27 This observation has been documented in other studies based on real-effort tournaments. A recent example is Freeman and Gelber (2010), where performance is not substantially altered in their single prize tournament when competitor’s past performance is revealed.

28 If an additional quadratic interaction between treatment K and the ability measure “Pretest” is included then there is a significant positive effect for subjects with medium ability. However, this effect is mainly driven by the one subject of highest ability (Pretest = 11) whose tournament performance is drastically lower than his trial performance. When this subject is eliminated from the sample, results of the quadratic specification are no longer significant. It should also be noted that the sample size is relatively small such that the power of testing the relationship between K and tournament performance is relatively low.

29 “Grade” is not significant for non-experienced subjects.

30 Again “Grade” is not significant in any specification for non-experienced subjects.
The inclusion of the interaction term implies that a representative subject of the base group in treatment K has low ability (zero correct sudokus in the five minute trials).\textsuperscript{31}

Results for experienced subjects are presented in Table 5. Taken together, results from REG (5) and REG (7) imply that AA affects subjects’ performance differently, depending on subjects’ ability levels, since coefficients in REG (7) become significant when the interaction between AA and Pretest is taken into account. While the coefficient for AA is large, positive, and significant at the 5\% level in this

\textsuperscript{31} The statistical effect of AA on subjects with higher ability can therefore be calculated as “AA” + “AA-Pretest”.

Table 3

Tournament performance (average no. of correctly solved sudokus with standard deviations in parenthesis) by treatment and school year.

<table>
<thead>
<tr>
<th></th>
<th>4th Grade</th>
<th></th>
<th>6th Grade</th>
<th></th>
<th>Overall</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E</td>
<td>NE</td>
<td>E</td>
<td>NE</td>
<td>E</td>
<td>NE</td>
</tr>
<tr>
<td>NK</td>
<td>28 (15.43)</td>
<td>16.08 (8.01)</td>
<td>38.93 (16.10)</td>
<td>24.67 (15.44)</td>
<td>31.27 (16.44)</td>
<td>20.38 (12.80)</td>
</tr>
<tr>
<td>K</td>
<td>29.88 (12.47)</td>
<td>17.69 (10.74)</td>
<td>43 (17.98)</td>
<td>29.09 (13.43)</td>
<td>35.37 (16.22)</td>
<td>22.92 (13.11)</td>
</tr>
<tr>
<td>AA</td>
<td>29.38 (13.78)</td>
<td>19.26 (9.48)</td>
<td>45.67 (12.04)</td>
<td>28.08 (12.12)</td>
<td>36.85 (15.31)</td>
<td>24.04 (11.80)</td>
</tr>
<tr>
<td>LH</td>
<td>27.59 (13.13)</td>
<td>23.36 (9.19)</td>
<td>44.86 (11.51)</td>
<td>29.50 (14.43)</td>
<td>35.39 (15.02)</td>
<td>27 (12.73)</td>
</tr>
<tr>
<td>LL</td>
<td>27.59 (12.26)</td>
<td>19.42 (11.79)</td>
<td>51.54 (11.44)</td>
<td>26 (9.01)</td>
<td>37.97 (16.82)</td>
<td>22.57 (10.85)</td>
</tr>
<tr>
<td>PH</td>
<td>29.67 (12.38)</td>
<td>17.93 (9.05)</td>
<td>46.92 (11.09)</td>
<td>26.54 (11.16)</td>
<td>37.68 (14.52)</td>
<td>22.07 (10.84)</td>
</tr>
<tr>
<td>PL</td>
<td>32.94 (17.36)</td>
<td>17.07 (7.60)</td>
<td>40.27 (12.38)</td>
<td>29.16 (12.83)</td>
<td>36.48 (15.37)</td>
<td>24.25 (12.42)</td>
</tr>
</tbody>
</table>

Fig. 3 (continued).
specification, the respective interaction variable “AA\*Pretest” is negative and also significant. These results suggest that experienced subjects in the base group (with low ability) solve on average 9.14 more sudokus when they compete with subjects favored by affirmative action policies. However, the negative and significant coefficient of the interaction term indicates that the higher the ability of experienced subjects (measured by “Pretest”), the lower the increase in performance under AA. Notice that these results imply that the performance of experienced subjects of highest ability under AA is lower than the performance of experienced subjects in the baseline treatment.

REGs (6) and (8) allow us to analyze whether the established results for the pooled AA treatment are driven by specific AA policies. REG (6) suggests that policy PH has a positive and significant effect on performance. If the interaction of our proxy for ability and the different AA treatments is also taken into account then REG (8) indicates that the AA treatments where the affirmative action policy is strong, i.e., treatments LH and PH, have significant positive effects on performance for low ability levels. Moreover, and in line with the results obtained from the pooled AA treatment in REG (7), the coefficients of the interaction term with respect to the strong AA treatments (“LH\*Pretest” and “PH\*Pretest”) are negative and significant at the 5% level. This suggests that the strong LH and PH treatments are the main contributors for the observed incentive effects. We thus conclude:

Result 2. We find no evidence that the implementation of Affirmative Action policies had a significantly negative (or positive) impact on average tournament performance of experienced subjects. However, there is some evidence that AA enhanced the performance of experienced subjects with lower ability levels. This performance enhancement declined for experienced subjects with higher ability levels. Most of this effect can be attributed to treatments where the AA policy was strong, i.e., PH and LH.

We now focus on non-experienced subjects. Results from the same specifications used for the experienced subject pool are presented in Table 6. REGs (9) and (10) reveal that AA has a positive and significant effect on tournament performance. Non-experienced subjects solve on average 3.98 more sudokus when they are favored by an affirmative action policy. REG (10) suggests that the incentive enhancing effects of affirmative action cannot be attributed to a specific type of policy because

If the dependent variable in the regressions from Table 5 is substituted with the total number of (correctly and wrongly) solved sudokus then results are very similar in all four regressions. These results are in line with the observation that for experienced subjects the number of wrongly solved sudokus is mostly not affected by AA (as it is shown in alternative specification where the number of wrongly solved sudokus is the respective dependent variable).
coefficients for all AA treatments (LH, LL, PH and PL) are similar in size and significance levels. Hence, there seems to be a performance enhancing effect of affirmative action policies on non-experienced subjects which is independent of its form or size. The lack of sensitivity to the specific type and size could be the result of non-experienced subjects’ lack of familiarity with the task, which may reduce their capacity to assess the relative size of compensations. The interaction term between AA and Pretest is included in REG (9) for the pooled AA treatments and in REG (10) for each AA treatment separately. In contrast to the group of experienced subjects there is no indication of different performance effects by low or high ability subjects in the different AA treatments because the interaction terms (but also the AA treatment effects) become mostly insignificant.34 This might be attributed to the fact that non-experienced subjects have less ground to assess their relative ability, since the task is new to them. An alternative explanation for the non-significant results in REG (11) and (12) might be the limited sample size for each treatment in combination with large heterogeneity in performance of subjects with different abilities.35 We now summarize the results for non-experienced subjects.

**Result 3.** Affirmative Action policies enhanced the performance of non-experienced subjects. We do not find evidence that this enhanced performance depends on the two different sizes or types of the implemented policy.

Finally, one concern about our experimental design is children’s capacity to understand the different incentives underlying the different

---

34 We again find significant results for non-experienced subjects when AA is interacted with the Pretest variable in a quadratic way. In particular, while performance is higher for subjects of all ability levels in the AA treatment, their performance is lower the higher their ability level. However, this result is once again driven by the one non-experienced subject in the K treatment with unexpected low tournament performance. When eliminating this subject from the sample, coefficients of the quadratic regression are no longer significant which is in line with the results from REG (11) and (12).

35 If the dependent variable in REG (11) and (12) is substituted with the total number of (correctly and wrongly) solved sudokus then the regression result show similar sized coefficients which are then mostly significant. These results suggest that non-experienced subjects tried to solve sudokus more carefully and diligently under AA, especially those subjects with low ability. Regression results where the dependent variable is the number of wrongly solved sudokus confirm this interpretation in the sense that AA had a significant negative effect on the number of wrongly solved sudokus.
treatments. In order to study this concern, we repeat the analysis separately for subjects from the 4th and 6th grade without reporting the specific details here. Qualitative results remain similar in the sense that the signs of the policy coefficients are mostly maintained among the age groups, but size and significance of the treatment effects is sometimes reduced. Results are actually much stronger for subjects from 4th grade than for 6th grade. For instance, experienced subjects from 4th grade react (in the specification without interaction effects) significantly negatively to knowledge about the asymmetry, while subjects from 6th grade do not. Similarly, the coefficients for AA have the same signs for both age groups but are substantially larger for subjects from 4th grade such that AA coefficients remain significant only for this age group.

For non-experienced subjects, knowledge does not affect behavior for any age group which confirms Result 1. The differences in performance with respect to the AA treatments between the two age groups of non-experienced subjects are similar to the difference observed for experienced subjects: While the signs of the respective AA coefficients are the same in the two age groups their sizes are substantially larger for younger subjects. This also affects significance levels in the sense that AA coefficients remain significant only for subjects from 4th grade.

Overall it seems that different age groups in the two subject pools react in a similar way (direction) to the implemented policies where these reactions are stronger for younger subjects. One potential explanation for the observed difference in behavior of different age groups might be different perceptions in the underlying incentive structure in the following sense: for older subjects the competitive pressure from the pure tournament (against a rival from a different school) might be sufficiently important to induce nearly maximal effort exertion. Hence, their reaction to the implementation of AA has limited scope. However, for younger subjects the incentive effects from the pure tournament are less important such that AA policies are more relevant for these subjects and, therefore, have a more substantial effect on performance in the younger age group.

3.4. The effects of affirmative action on beliefs

The implementation of AA might not only have a direct effect on performance but it could also affect subjects’ expectations about winning their respective tournament. This issue is of importance since it allows us to study whether affirmative action undermined self-confidence of the respective subjects. We use question 6 of the post-experiment questionnaire in which subjects ranked their expectation of winning the tournament against their respective rival on an ordinal scale from 1 (“Definitely Not”) to 5 (“Definitely”). As there was no information about the identity and characteristics of the respective opponent we use these answers as a measure of confidence in winning. Notice that this measure is obtained in a post-experiment questionnaire, and thus it may already incorporate information not only about how beliefs are affected by the treatments but also by the individual performance in the respective tournament.

Our analysis follows the same structure as the two previous subsections. Table 7 shows the results of OLS regressions comparing the NK and K treatments, while Table 8 shows the respective results of OLS regressions comparing the AA treatments with the K treatment. We report robust standard errors, clustered by treatment and class. The dependent variable is now the answer to question no. 6, that is, the score assigned by subjects to their belief in winning the respective tournament.36 Results for experienced subjects are

---

36 As the dependent variable has an ordered rank structure, an alternative estimation strategy would be to estimate treatment effects by ordered probit regressions and then to analyze estimated average partial effects. The results of this alternative estimation strategy are very similar to the results presented here.

---

Table 4

<table>
<thead>
<tr>
<th></th>
<th>Experienced (Pretest)</th>
<th>Experienced (Posttest)</th>
<th>Non-experienced (Pretest)</th>
<th>Non-experienced (Posttest)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>−6.98**</td>
<td>−9.79</td>
<td>1.40</td>
<td>1.16</td>
</tr>
<tr>
<td>Pretest</td>
<td>−1.28</td>
<td>−2.28</td>
<td>0.76</td>
<td>2.47</td>
</tr>
<tr>
<td>Grade</td>
<td>2.33**</td>
<td>2.67**</td>
<td>2.10</td>
<td>2.67</td>
</tr>
<tr>
<td># Observations</td>
<td>59</td>
<td>60</td>
<td>49</td>
<td>47</td>
</tr>
</tbody>
</table>

Notes: The dependent variable in all regressions is individual tournament performance, i.e. the number of correctly solved sudokus in the respective tournaments. "K" is the tournament dummy for treatment K, "Pretest" denotes performance in the pretest, i.e. the no. of correctly solved sudokus in the trial (minimum is 0 for the E and NE group, maximum is 6 for the E group and 12 for the NE group). "Grade" is a normalized average of individual school grades (minimum is 0, maximum is 5). Year is a school year dummy (4th year is 0, 6th year is 1). "Gender" is a gender dummy (male is 0, female = 1). All estimations are the result of Ordinary Least Square regressions. Robust standard errors, clustered by treatment and class are in parentheses, where ** denotes significance at the ** level, *** denotes significance at the 1% level.

---

Table 5

<table>
<thead>
<tr>
<th></th>
<th>REG (5)</th>
<th>REG (6)</th>
<th>REG (7)</th>
<th>REG (8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>−7.63***</td>
<td>−8.33**</td>
<td>−14.22***</td>
<td>−14.47***</td>
</tr>
<tr>
<td>Pretest</td>
<td>(3.06)</td>
<td>(3.05)</td>
<td>(3.74)</td>
<td>(3.84)</td>
</tr>
<tr>
<td>AA</td>
<td>0.88</td>
<td>9.14**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AA-Pretest</td>
<td>(1.59)</td>
<td>(3.77)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LH</td>
<td>0.80</td>
<td>(0.78)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>(0.92)</td>
<td>(0.87)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>0.84</td>
<td>(0.85)</td>
<td></td>
<td></td>
</tr>
<tr>
<td># Observations</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>150</td>
</tr>
</tbody>
</table>

Notes: The dependent variable in all regressions is individual tournament performance of experienced subjects, i.e. the no. of correctly solved sudokus in the respective tournament. "AA", “LH”, “LL”, “PH”, “PL” are the treatment dummies for the respective treatments. "Pretest" denotes performance in the pretest, i.e. the no. of correctly solved sudokus in the trial (minimum is 0 for the E and NE group, maximum is 6 for the E group and 12 for the NE group). "Gender" is a gender dummy (male is 0, female = 1). All estimations are the result of Ordinary Least Square regressions. Robust standard errors, clustered by treatment and class are in parentheses, where ** denotes significance at the ** level, *** denotes significance at the 1% level.
by treatment and class are in parentheses, where grades (minimum is 0, maximum is 5). Year is a school year dummy (4th year is 0, 6th year is 1). All estimations are the result of Ordinary Least Square regressions. Robust standard errors, clustered by class are in parentheses, and in REG (11) and (12) as well as in REG (19) and (20).

We find that in general, “Pretest” has a positive (and in most cases significant) coefficient in all regressions, indicating that subjects with higher ability believe that they are going to win the tournament with higher probability. However, the knowledge of the differences in experience does not seem to affect the beliefs of the subjects to a large extent: Table 7 reveals that the coefficient of the K treatment is mostly non-significant with one exception in REG (13). More importantly, the implementation of AA does not significantly affect reported beliefs of experienced subjects as it is shown in Table 8, REG (17) and (18). For non-experienced subjects the treatment effect is positive but small and only significant for the interaction effect with the ability measure “Pretest” in REG (20).

These results suggest that experienced subjects do not feel frustrated by the introduction of AA which is also consistent with their respective fairness perception of the AA policies; the analysis of question no. 8 of the post-experimental questionnaire reveals that experienced subjects perceive AA policies as significantly less fair than the benchmark treatment K (non-parametric MW-tests between separate and joint AA treatments are highly significant with the exception of the PL treatment). In some sense experienced subjects are clearly aware of the disadvantageous implications of affirmative action policies on themselves but seem to be willing to compensate the bias induced by AA through exerting more effort such that this effect is neutralized with respect to the perceived probability of winning.

Similar observations can be made with respect to non-experienced subjects: their beliefs of winning the tournament are neither positively nor negatively affected by AA policies. However, the analysis of fairness perceptions reveals that treatments are perceived differently by subjects. Treatments where AA is implemented are perceived differently by subjects. Treatments where AA is implemented are perceived differently by subjects. Treatments where AA is implemented are perceived differently by subjects. Treatments where AA is implemented are perceived differently by subjects. Treatments where AA is implemented are perceived differently by subjects. Treatments where AA is implemented are perceived differently by subjects. Treatments where AA is implemented are perceived differently by subjects. Treatments where AA is implemented are perceived differently by subjects. Treatments where AA is implemented are perceived differently by subjects.

3.5. The effects of affirmative action on the selection of tournament winners

An important concern raised in the context of affirmative action policies is that the selected pool of candidates may be of lower ability because of the higher proportion of selected disadvantageous individuals who may perform poorly. There are two different approaches to answer this question, crucially depending on the

Notes: The dependent variable in all regressions is individual tournament performance of non-experienced subjects, i.e. the no. of correctly solved sudokus in the respective tournament. “AA”, “LH”, “LL”, “PH”, “PL” are the treatment dummies for the respective treatments. “Pretest” denotes performance in the pretest, i.e. the no. of correctly solved sudokus in the trial (minimum is 0 for the E and NE group, maximum is 6 for the E group and 12 for the NE group). “Grade” is a normalized average of individual school grades (minimum is 0, maximum is 5). Year is a school year dummy (4th year is 0, 6th year is 1). “Gender” is a gender dummy (male is 0, female=1). All estimations are the result of Ordinary Least Square regressions. Robust standard errors, clustered by class are in parentheses, where *** denotes significance at the 1% level, ** denotes significance at the 5% level and * denotes significance at the 1% level. (a) For one non-experienced subject “Grade” was not available. Another subject arrived late and did not participate in the practice rounds. Such observations are omitted.

presented in REG (13) and (14) as well as REG (17) and (18), while results for non-experienced subjects are presented in REG (15) and (16) as well as in REG (19) and (20).

The potential caveat of low sample size also holds for the specifications in Tables 7 and 8.
objective of these policies. First, the objective may be to select individuals according to unobserved ability, which is assumed to be positively correlated with performance then selecting a similar proportion of the best performing individuals in both groups should lead to a pool of selected individuals that consists of the highest ability individuals overall. Table 9 reports the percentage of tournaments winners from the NE school as a result of performing all possible pair-wise matches between competitors from both schools in each treatment. The percentage of NE tournament winners in the 4th grade (58.29%) is higher than in the 6th grade (45.81%), suggesting that the sizes of the compensations may have been a bit too high (low) given the performance of 4th (6th) graders. Comparing the

Note: The dependent variable in all regressions is the individual belief in winning the respective tournament, i.e., the answer to question 6 of the questionnaire, where answer “Definitely Not” is codified as 1 (minimum) and “Definitely” as 5 (maximum). “AA”, “PH”, “PH”, “PH”, “PH” are the treatment dummies for the respective treatments. “Pretest” denotes performance in the pretest, i.e. the no. of correctly solved sudokus in the trial (minimum is 0 for the E and NE group, maximum is 6 for the E group and 12 for the NE group). “Grade” is a normalized average of individual school grades (minimum is 0, maximum is 5). “Gender” is a gender dummy (male is 0, female = 1). All estimations are the result of Ordinary Least Square regressions. Robust standard errors, clustered by class are in parentheses, where ** denotes significance at the 1% level, *** denotes significance at the 5% level and * denotes significance at the 1% level.

Table 8
Expected winning probability in treatment K vs. AA.

<table>
<thead>
<tr>
<th></th>
<th>Experienced</th>
<th>Non-experienced</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>REG (17)</td>
<td>REG (18)</td>
</tr>
<tr>
<td>Constant</td>
<td>3.12***</td>
<td>(0.41)</td>
</tr>
<tr>
<td></td>
<td>2.76***</td>
<td>(0.44)</td>
</tr>
<tr>
<td></td>
<td>2.34***</td>
<td>(0.34)</td>
</tr>
<tr>
<td></td>
<td>2.61***</td>
<td>(0.31)</td>
</tr>
<tr>
<td>AA</td>
<td>0.12</td>
<td>(0.31)</td>
</tr>
<tr>
<td></td>
<td>0.18</td>
<td>(0.34)</td>
</tr>
<tr>
<td></td>
<td>0.24</td>
<td>(0.24)</td>
</tr>
<tr>
<td></td>
<td>−0.11</td>
<td>(0.11)</td>
</tr>
<tr>
<td>AA+Pretest</td>
<td>0.10</td>
<td>(0.08)</td>
</tr>
<tr>
<td>Pretest</td>
<td>0.16***</td>
<td>(0.05)</td>
</tr>
<tr>
<td></td>
<td>0.24***</td>
<td>(0.05)</td>
</tr>
<tr>
<td>Grade</td>
<td>−0.15</td>
<td>(0.09)</td>
</tr>
<tr>
<td></td>
<td>−0.15</td>
<td>(0.09)</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>(0.08)</td>
</tr>
<tr>
<td></td>
<td>0.09</td>
<td>(0.08)</td>
</tr>
<tr>
<td>Year</td>
<td>0.07</td>
<td>(0.18)</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>(0.14)</td>
</tr>
<tr>
<td>Gender</td>
<td>−0.02</td>
<td>(0.19)</td>
</tr>
<tr>
<td></td>
<td>−0.02</td>
<td>(0.19)</td>
</tr>
<tr>
<td># Observations</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td>Pseudo R²</td>
<td>0.10</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Notes: The percentage of non-experienced tournament winners by treatment and school year.

Table 9
Percentage of non-experienced tournament winners by treatment and school year.

<table>
<thead>
<tr>
<th></th>
<th>4th year</th>
<th>6th year</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>NK</td>
<td>25</td>
<td>23.81</td>
<td>24.42</td>
</tr>
<tr>
<td>K</td>
<td>21.27</td>
<td>27.27</td>
<td>23.94</td>
</tr>
<tr>
<td>AA</td>
<td>58.29</td>
<td>45.81</td>
<td>51.81</td>
</tr>
<tr>
<td>LH</td>
<td>83.42</td>
<td>57.14</td>
<td>72.32</td>
</tr>
<tr>
<td>LL</td>
<td>49.51</td>
<td>10.49</td>
<td>31.84</td>
</tr>
<tr>
<td>PH</td>
<td>61.43</td>
<td>55.03</td>
<td>58.40</td>
</tr>
<tr>
<td>PL</td>
<td>40.27</td>
<td>53.68</td>
<td>45.96</td>
</tr>
</tbody>
</table>

Table 10
Average tournament performance (standard deviations in parenthesis) by all possible tournament winners in each treatment and school year.

<table>
<thead>
<tr>
<th></th>
<th>4th year</th>
<th>6th year</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>NK</td>
<td>30 (13.54)</td>
<td>42.04 (15.52)</td>
<td>35.81 (15.71)</td>
</tr>
<tr>
<td>K</td>
<td>31.83 (11.62)</td>
<td>46.91 (13.03)</td>
<td>37.75 (14.23)</td>
</tr>
<tr>
<td>AA</td>
<td>29.70 (13.35)</td>
<td>43.36 (12.53)</td>
<td>36.53 (14.63)</td>
</tr>
<tr>
<td>LH</td>
<td>28.20 (11.98)</td>
<td>42.91 (12.58)</td>
<td>36.22 (14.31)</td>
</tr>
<tr>
<td>LL</td>
<td>29.73 (12.79)</td>
<td>51.09 (11.12)</td>
<td>38.54 (16.04)</td>
</tr>
<tr>
<td>PH</td>
<td>27.94 (11.00)</td>
<td>41.30 (12.80)</td>
<td>33.90 (13.56)</td>
</tr>
<tr>
<td>PL</td>
<td>32.56 (16.27)</td>
<td>41.08 (11.51)</td>
<td>37.36 (14.41)</td>
</tr>
</tbody>
</table>

Notes: *** denotes significance at the 1% level, **** denotes significance at the 5% level and ***** denotes significance at the 1% level.

Table 11
Percentage change of the average correct sudokus by tournament winners.

<table>
<thead>
<tr>
<th></th>
<th>4th Year</th>
<th>6th Year</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>NK vs. K</td>
<td>−5.76**</td>
<td>−10.38***</td>
<td>−5.14**</td>
</tr>
<tr>
<td>NK vs. AA</td>
<td>1.02</td>
<td>−3.05**</td>
<td>−1.95*</td>
</tr>
<tr>
<td>K vs. AA</td>
<td>7.19**</td>
<td>8.17***</td>
<td>3.36</td>
</tr>
</tbody>
</table>

Notes: *** denotes significance at the 1% level, **** denotes significance at the 5% level and ***** denotes significance at the 1% level.
significant loss in each school year when AA was implemented (a loss of 7.19% for 4th year and a of 8.17% for 6th year). This indicates that there was a loss in average performance although it was not large.\footnote{Notice that performance is a significantly lower 7% when comparing the performance by tournament winners under AA with respect to the hypothetical situation in which affirmative action was announced (such that the performance inducing effects are present) but then not used to select tournament winners. This type of policy, although more efficient from a performance point of view, may be difficult to justify and not credible in the long term.}

Thus we conclude:

**Result 4.** While affirmative action policies induced a leveled playing field, they did not do so at the expense of a large loss in average performance by subjects selected as tournament winners.

### 4. Conclusion

In this experimental study we exploit a situation where an exogenous disadvantage among two otherwise similar subject pools exists. This allows us to analyze the incentive effects of affirmative action in a real effort tournament setting where subjects have systematic differences in capacities to compete with respect to the relevant task. Hence, the implementation of affirmative action can be evaluated without potential confounding effects due to self-selection, complex behavioral aspects like stereotype threats, or the potential artificiality of a laboratory setting. Moreover, our experimental design allows us to derive a measure for individual ability in the relevant tournament task which can be used to control for individual heterogeneity in task abilities and therefore isolate the incentive effects of the implementation of affirmative action.

Our empirical analysis of the experimental results suggests that the implementation of affirmative action does not lead to negative incentive effects on tournament performance. If at all, there is rather evidence that AA results in enhanced performance by a large fraction of participants and a balanced pool of selected winners, while there is only a small decline in the average performance of selected winners.

Our study has some limitations. First, as in real-life situations where affirmative action policies are actually applied, the appropriate size of the affirmative action bias cannot be precisely specified ex-ante because the precise extent of the disadvantage is unknown to the tournament designer. Second, our set up is somehow artificial since AA is usually implemented in situations where not only the size of the disadvantage between groups is unknown but also the nature of this difference is more ambiguous and complex, for example, due to racial or ethnic differences, stereotype threats or different propensities to compete across different genders. However, our design allows us to identify the incentive effects of affirmative action resulting from the reduction of an existing asymmetry in actual capacities to compete, thereby abstracting from the intricate nature of discrimination in actual applications. Based on this set-up our study suggests that reducing identified asymmetries to compete may actually improve performance. Third, experienced subjects in our tournaments are better informed of their own abilities than non-experienced subjects, and thus reactions to affirmative action policies by both types of subjects may be asymmetric due to the effort choice being less risky for more informed subjects. Moreover, our experiment is done with children, whose reactions to affirmative action policies may differ from other populations due to possible differences in understanding the underlying incentives behind these policies. Finally, our design only allows us to study performance effects which are contemporaneous to the implemented policy, while long term effects are potentially important in this setting. We plan to overcome these limitations in future research.

### Appendix A. Experimental instructions

Below a translation of the experimental instructions used in the experiment is provided. Instructions for all treatments and schools were identical apart from the changes here indicated. Sentences in bold were not included in the NK treatment. The sentences in bold and italics were only included in the treatments with affirmative action (LH, LL, PH, PL). Words in (parentheses) indicate changes between the schools with experienced and non-experienced subjects and changes in the type of the AA policies (lump-sum or proportional). Sizes of the AA policies varied as explained in Section 2. Numerical examples varied in order to reflect changes in sizes, but were created such as the results of both contestants were the same. A whole set of instructions is available upon request. Instructions were originally written in Spanish.

**Pre-instructions**

Your Code: __________________________________________

Thank you for participating. First, we are going to explain what you will be doing.

You have to fill in grids with the numbers 1, 2, 3 and 4.

To do this you have to use the following rules:

1. All boxes in a grid must be filled in with a number.
2. The same number can appear only once in each column (vertical).
3. The same number can appear only once in each row (horizontal).
4. The same number can appear only once in each square. Each grid is divided in 4 squares, marked in bold lines.
5. In each grid all numbers 1, 2, 3 and 4 must be in each column, each row, and each square.
Here are some examples:

This row is **completed wrongly**
because the 4 appears twice (rule3)

This square is **completed wrongly**
because the 1 appears twice (rule4)

This column is **completed wrongly**
because the 3 appears twice (rule2)

This column is **completed correctly.**

This row is **completed correctly.**

This row is **completed correctly.**

This square is **completed correctly.**

This is an example of a correctly completed grid.
Before starting you have 5 min to complete the following grids to check whether you have understood the rules. We will give you the correctly completed grids after the 5 min period.

Please remain silent and on your seat without disturbing anyone during the whole practice. Raise your hand after you have finished all grids and we will pick them up. Good luck!

Instructions

You are randomly matched with another student, your matched participant, from another school similar to yours, who is completing the same grids as you are.

The students at the other school have (NOT) learned before how to solve these types of grids because it was (NOT) taught to them in their math classes.

You have now 30 min time to complete as many grids as possible with the numbers 1, 2, 3 and 4 on the formulaires that we are now going to distribute. We will compare how many grids you have solved correctly with the number of correctly solved grids by your matched participant from the other school:

- If you have correctly solved more grids then you will earn a 7 EU voucher that you can redeem in “La Casa del Libro”, where you can buy books, collector’s cards, toys, music or comics.
- If you have correctly solved less grids then you will not earn the voucher.
- If you have correctly solved the same number of grids, then a toss of a coin will be used to determine who earns the voucher.

To compensate (the other students) for the fact that (they)/(you) have (less)/(more) practice (than you) we are going to give (them)/(you) (20 extra grids)/(1 grid more for each grid that (they)/(you) solve correctly).

For example (example provided for the PH Treatment):

- If your matched participant correctly solves 12 grids, they count as 12 + 12 = 24 grids. Therefore you will earn the voucher if you solve correctly 25 grids or more.
- If your matched participant correctly solves 30 grids, they count as 30 + 30 = 60 grids. Therefore you will not earn the voucher if you solve correctly 59 grids or less.
- If your matched participant correctly solves 20 grids, they count as 20 + 20 = 40 grids. Therefore, if you solve correctly 40 grids, a toss of a coin determines whether you earn the voucher.
The numbers of this example are chosen randomly and do not indicate how many grids a student can solve correctly. We would like to inform you that we have studied the results of other students of your age from other schools who completed the same grids: The maximum number of grids that somebody managed to solve correctly in 30 min were 81 grids and the minimum was 0 grids. On average the students completed around 25 grids correctly. Remember that only correctly solved grids count. Wait to turn the answer sheet until we tell you to do so. You have 30 min. Good luck!

Thank you for your participation.

Your Code:

Final Questionnaire

Please answer the following questions:

1. How did you find today’s task?
   Interesting   Entertaining   A bit long   Boring
2. How many grids like these have you tried before?
   None   Between 1 and 5   Between 6 and 20   Between 20 and 40   More than 40
3. If you have tried solving grids like these before, where did you do it? _______________________
4. How many grids do you think you have solved correctly today? _______________________
5. How many grids do you think your partner of the other school has solved correctly? _______________________
6. Do you think you are going to get the voucher?
   Definitely   Probably yes   I don’t know   Probably not   Definitely not
7. Do you think it was a good idea to compensate the students of the other school that did not do grids like this before in school?
   YES   NO
8. The competition with the students of the other school from my perspective seemed to be:
   Fair   Rather Fair   A bit Unfair   Unfair   Rather Unfair   Very Unfair
9. Any other comment? __________________________________________________

References