NBER WORKING PAPER SERIES

AFFIRMATIVE ACTION AND HUMAN CAPITAL INVESTMENT: EVIDENCE FROM A RANDOMIZED FIELD EXPERIMENT

Christopher Cotton
Brent R. Hickman
Joseph P. Price

Working Paper 20397
http://www.nber.org/papers/w20397

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2014

We would like to acknowledge helpful comments from John List and participants of seminars at the University of Chicago, the 2013 ASSA meetings, the 2013 North American Summer Meetings of the Econometric Society, and Brigham Young University. We also wish to acknowledge the outstanding assistance of Joe Patten, who played a crucial role in executing this research. Funds for the research project came from author university research funds, a James W. McLamore Research Award from the University of Miami, and grant from the Spencer Foundation. The experiment was conducted when Cotton was a faculty member at the University of Miami. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Affirmative Action and Human Capital Investment: Evidence from a Randomized Field Experiment
Christopher Cotton, Brent R. Hickman, and Joseph P. Price
NBER Working Paper No. 20397
August 2014
JEL No. C93,D44,D82,J15,J24

ABSTRACT

The empirical literature on Affirmative Action (AA) in college admissions tends to ignore the effects admissions policies have on incentives of students to invest developing pre-college human capital. We explore the incentive effects of AA using a field experiment that creates a microcosm of the college admissions market. Our experimental design is based on the asymmetric, multi-object, all-pay auction framework in Bodoh-Creed and Hickman (2014). We pay 5th through 8th grade students based on their performance on a national mathematics exam relative to other competitor students, and observe the use of a study website as students prepare for the exam. An AA treatment favors "disadvantaged" students by reserving prizes for lower grade students who on average have less mathematics training and practice. We find that the AA policy significantly increases both average time investment and subsequent math achievement scores for disadvantaged students. At the same time, we find no evidence that it weakens average human capital investment incentives for advantaged students. We also find strong evidence that AA can narrow achievement gaps while promoting greater equality of market outcomes.

Christopher Cotton
Queen's University
Department of Economics
Kingston, Ontario K7L 3N6
cotton@econ.queensu.ca

Brent R. Hickman
Department of Economics
University of Chicago
1126 E. 59th Street
Chicago, IL 60637
hickmanbr@uchicago.edu

Joseph P. Price
Department of Economics
Brigham Young University
162 FOB
Provo, UT 84602
and NBER
joseph_price@byu.edu

1. Introduction

Affirmative Action (AA) is the practice of granting preferential treatment to underrepresented (UR) demographic groups when allocating contractual, employment, or educational opportunities. It was first mandated by the Kennedy Administration in the 1960s, and has since been widely implemented in public procurement, education and hiring. Today, AA is a pervasive fixture of American college admissions; though it has generated much controversy.[1] AA has also been widely implemented outside the United States, from Malaysia to Northern Ireland, and in India where *Reservation Law*, a set of strict racial and ethnic quotas, is imposed by constitutional edict.

In the United States the rationale for AA is that the college admissions process is effectively a competition, where black and Hispanic children—attending lower quality schools, being less affluent, and having parents who are less educated— are at an inherent disadvantage to whites and Asians due to the lingering effects of past institutionalized racism.[2] In turn, AA seeks to help compensate for the disadvantage that UR minorities face by giving special consideration for race in allocation of college seats. There is a substantial empirical literature studying the allocative and welfare effects of AA in post-secondary admissions. Most of it focuses on the direct impact of an AA policy on student and school outcomes starting at the time of admissions.[3]

However, the literature to a large extent has ignored the impact that an AA policy may have on student behavior *prior* to college admissions, as they study, participate in extra curricular activities, and otherwise use time inputs to build human capital before sending out their first college application. Because AA policies alter what level of academic performance is required to get into different colleges, important questions remain if high school students are forward looking: to what extent does it shape effort incentives? Does

[1]The US Supreme Court has deliberated on the legality of racial considerations in college admissions at least five times, including the cases of *Schuette v. Coalition to Defend Affirmative Action* (2014), *Fisher v. Texas* (2013), *Grutter v. Bollinger* (2003), *Gratz v. Bollinger* (2003), and *Regents of the University of California v. Bakke* (1978). At the time of this writing, *Fisher v. Texas* (2013) had been remanded back to the US 5th Circuit Court and was in the appeals process.

[2]Lyndon B. Johnson, Kennedy's successor, was the first American president to implement AA. In his 1965 commencement address at Howard University, Johnson articulated this idea as a motivation for AA: "*You do not take a person who, for years, has been hobbled by chains and liberate him, bring him up to the starting line of a race and then say, 'you are free to compete with all the others,' and still justly believe that you have been completely fair. Thus it is not enough just to open the gates of opportunity. All our citizens must have the ability to walk through those gates... To this end equal opportunity is essential, but not enough, not enough. Men and women of all races are born with the same range of abilities. But ability is not just the product of birth. Ability is stretched or stunted by the family that you live with, and the neighborhood you live in–by the school you go to and the poverty or the richness of your surroundings. It is the product of a hundred unseen forces playing upon the little infant, the child, and finally the man.*"

[3] Bowen and Bok [1998], Arcidiacono [2005] and Howell [2010], have attempted estimation of counterfactual racial admissions profiles in a color-blind world. Loury and Garman [1995], Sander [2004], Long [2008], Rothstein and Yoon [2008], and Chambers, Clydesdale, Kidder, and Lempert [2005], have estimated the impact of AA on graduation rates among UR minority groups.

a higher level of diversity on college campuses come at a cost in terms of human capital investment during high school? Does AA decrease incentives for UR minorities (its beneficiaries) or is the opposite true? Does it impact incentives for non-minorities? In turn, how does AA shape the racial achievement gap, if at all?

In this paper we develop a field experiment designed to investigate how an AA policy can change a student's tradeoff between labor (*i.e.*, investment of time into human capital accumulation) and leisure, as well as subsequent exam performance. Our experiment is designed to mirror the theoretical model of Bodoh-Creed and Hickman [2014], who apply an asymmetric, multi-object, all-pay auction framework to study human capital investment in a high school/college admissions context. We present an abbreviated version of their model in Section 3, and outline theoretical predictions of how the AA policy in our study will alter incentives by disadvantaged (*e.g.*, UR minorities) students and by advantaged students. Our analysis is designed to test these and other theoretical predictions in an attempt to shed light on the questions above.

Our experiment involves paying 5^{th} through 8^{th} grade students based on their relative performance on the American Mathematics Competition 8 (AMC8), a national mathematics exam. In order to provided a clean test of theory, we use age/grade cohort as our demographic delimiter. This distinction mirrors racial differences in that our disadvantaged students have, on average, less mathematics training and practice, while filtering out cultural differences which could confound the effects we seek to test.

Students are divided into two treatments for a math competition with real cash prizes, a "color-blind" control treatment and an AA treatment. In the control treatment, students compete against others in their own grade and an adjacent grade for prizes (we ran one contest for 5^{th} and 6^{th} graders and another for 7^{th} and 8^{th} graders), and test scores alone determine one's pay-off. In a second treatment, prizes are reserved for allocation to disadvantaged students, meaning that competition occurs only *within* each demographic group, while the distribution of prizes is left unchanged. This treatment represents a "quota" AA policy, an prize allocation rule which assures equal division of prizes across the disadvantaged and advantaged groups.

Each treatment started with a pre-exam based on the previous year AMC8, after which students were informed of their own score and the distribution of scores within their competition group. At this point, the students were told that their classes were going to take that year's AMC8 , and were given details about how their relative performance on the exam would determine their payment in the contest, and that they had 10 days to prepare for the exam. In order to track human capital investment during this period, we developed a website that provided AMC8 practice materials to help test subjects prepare for the exam. All students received access to this website during the 10-day

investment period between the pre-test and final exam, but they were left with the individual decision of whether and how much to use it on their own at home during free time. Meanwhile, our website monitored their individual choices by tracking who logged onto the site, how much time they spent, and the number of practice problems attempted.

We work with actual teachers to implement the competition in a classroom setting using materials similar to the curriculum students are regularly exposed to. Aside from the presence of short-term cash incentives for learning, this natural classroom and home learning environment created trade-offs familiar to test subjects' everyday experiences, and which mimic scenarios they will face when preparing for the competitive college admissions process during high school.

We find strong evidence that AA substantially boosted human capital investment activities by our disadvantaged group. We see large increases both on the extensive margin—the number of people willing to log on at least once—and on the intensive margin—as measured by time spent, math subject categories explored, and number of problems attempted. We also find that AA significantly improves disadvantaged students' exam scores, while significantly narrowing gaps between them and the advantaged group, on average. We also find little or no evidence that these gains come at the expense of less investment among advantaged students, on average; if anything our experimental data appear to favor a slight increase of investment instead. Finally, we find suggestive evidence that our test subjects as a group produced qualitative patterns of behavior consistent with the finer predictions of the theoretical predictions of Bodoh-Creed and Hickman [2014] by individual ability level within each demographic group.

The remainder of this paper has the following structure. Section 2 gives an overview of the previous literature and explains how our paper and contribution relate to what has been done before. Section 3 briefly outlines a special case of Bodoh-Creed and Hickman [2014]'s theoretical framework to illustrate model intuition and motivate our experimental design. Section 4 describes the structure of our field experiment in more detail. Section 5 presents and discusses our experimental results. Section 6 concludes, and an appendix expounds on some technical details of our study and presents additional tables and graphs.

2. Related Literature

There is a substantial empirical literature studying AA and its impact on college admissions. Bowen and Bok [1998] used student-level applications data to estimate the preference given to minority students by admissions officers at elite schools. Arcidiacono [2005] and Howell [2010] estimated structural models which adjust counterfactuals

for changes in minority application behavior induced by policy shifts. All three of these studies estimate that college placement without AA would be substantially less favorable to blacks. A related vein of the literature focuses on mismatching, or the idea that AA may cause black students to be placed higher but then graduate less frequently or from less lucrative majors, due to being unprepared for more academically demanding environments. Loury and Garman [1995] and Sander [2004] present evidence supporting mismatching. Other empirical work, such as Long [2008], Rothstein and Yoon [2008], and Chambers et al. [2005], suggest that some mismatching may occur, but its magnitude is relatively small and is outweighed by the benefits of placing blacks into higher quality institutions.

Throughout the literature on AA, a common thread is that SAT scores are used as a proxy for student ability, and assumed to be fixed with respect to variations in admissions criteria.[4] A new theory developed by Bodoh-Creed and Hickman [2014] argues that students may rationally adjust their pre-college academic effort in response to admissions policies. Therefore, SAT scores are a function of both ability and market incentives induced by AA. Therefore, assuming that SAT scores are independent of the prevailing college admissions policy may produce naive counterfactuals which are subject to the Lucas critique. In this paper our objective is to test this theory in a field experimental study of how students adjust their effort (influencing their test scores) in response to changes in AA policy.

This paper contributes to a new literature recognizing that AA policies may change the incentives governing pre-college human capital accumulation. Hickman [2013] estimated a structural model based on Bodoh-Creed and Hickman [2014], and conducted a counterfactual analysis of admissions, investment, and welfare under alternative AA policies. He finds evidence that AA increases the stock of minority human capital, and shows that in comparing AA policies, an admissions quota performs better than preference-based AA (*e.g.*, score bonuses) in US college admissions. Ferman and Assuncao [2011] shows that Brazilian high-school students change their academic effort in response to the introduction of an admissions quota at elite universities in Brazil. The study finds that the admissions quota decreased incentives for the group favored by the quota to exert effort, which in turn increased the achievement gap between blacks and non blacks by 25%. While seemingly contradictory, the results in Hickman [2013] and Ferman and Assuncao [2011] are not necessarily inconsistent with each other. Bodoh-Creed and Hickman [2014]'s theoretical model shows that the introduction of an AA policy may decrease the effort of high ability minority students while it increases the

[4]In the US, the SAT, or "scholastic aptitude test," is the most common standardized college admissions examination.

effort of lower ability students. Whether the overall impact of an AA policy on minority student effort is positive or negative depends on the distribution of students, and the ability of those most effected by the policy.[5]

A small handful of experimental papers also test the link between AA and effort. Schotter and Weigelt [1992] replicated a two-player contest in a laboratory setting, where asymmetry was induced by assigning participants one of two exogenous cost functions. The participants wrote down how many units of "effort" to spend during the contest, which directly determined both their costs and win probabilities. Calsamiglia, Franke, and Rey-Biel [2013] conduct a related field experiment in which 10-13 year old children compete against each other solving Sudoku puzzles. Students differ in whether they were exposed to Sudoku as part of their school's mathematics curriculum.

Our paper builds upon these earlier experimental studies in a number of important ways. First, we include an investment period between assignment to a treatment group and our final exam, during which time we monitor student time usage at home in a non-invasive way. This is a novel feature of our study because it provides us with a window into individuals' labor-leisure trade-offs, rather than focusing solely on in-class effort during a task. Our experimental design allows us to assess the impact of AA on both study time and final performance outcomes. Second, mathematics preparation that enables higher performance on the AMC8 exam is closer to the preparation that enables higher performance on college entrance exams, and is therefore more readily interpretable as human capital investment. Third, we argue that our experimental competition more closely mirrors important aspects of actual college admissions markets: it is a multi-player contest (with participants numbering in the hundreds) in which students compete against everyone else in their treatments group for a set of homogeneous outcomes. Our format also allows for overlap in ability distributions across demographic groups, rather than assuming no overlap as previous experimental work has done. This allows us to compare the performance *distributions* of students from advantaged and disadvantaged demographic groups to see whether different qualitative effects at different quantiles appear, as predicted by Bodoh-Creed and Hickman [2014]. Fourth, by working with test subjects regular teachers, using materials that were already being offered, and allowing for study choices at home, we create a natural setting in which test subjects are making decisions similar to those which will lead to their ultimate college placement outcomes.

[5]Ferman and Assuncao [2011] estimate a negative overall effect of the policy change, which would have predominantly impacted high-ability students since it applied only to top universities in Rio De Janeiro. Likewise, Hickman [2013] estimated that AA negatively affects investment among a small mass of the highest ability minorities in the US, even though on average it is beneficial.

Finally, our analysis is also related to studies of performance pay in primary and secondary schools. Recent studies including Bettinger [2012], Leuven, Oosterbeek, and van der Klaauw [2010] and Fryer [2011] pay students based on their individual academic performance and find mixed results about whether paying students can improve student effort and outcomes.[6] Our experiment pays participants based on their performance *relative* to other students, and therefore involves financial incentives more similar to those in Kremer, Miguel, and Thornton [2009], which studies a program awarding merit scholarships based on relative student performance. Our analysis, like Kremer et al. [2009], finds significant evidence that student performance responds to financial incentives in a competitive environment.[7] Unlike most of these analyses, our experimental design allows us to directly monitor student study effort, and assess how it (rather than only final performance measures) responds to incentives. Additionally, our focus is significantly different from these other papers, as we are interested in assessing the impact of AA policies on performance, rather than the impact of pay for performance policies themselves.

3. Theoretical Background

We model our field experiment after the general framework developed in Bodoh-Creed and Hickman [2014], which frames the college admissions market in terms of a multi-object all-pay auction. For illustrative purposes, we present a special case of that framework below.

[6] Studying primary school students, Bettinger [2012] shows that pay for performance programs can significantly increase math performance, but no evidence that they increase performance in other subjects. Leuven et al. [2010] shows that merit pay programs may improve the performance of high ability students while decreasing the performance of low ability students. Fryer [2011] reports results from three experiments, where students are paid based on one of the following: number of books they read, exam performance, or final grades earned. Students in English speaking classrooms tend to perform better on reading assessments when they are paid to read books, while students in bilingual classrooms tend to decrease their performance given the same incentives. Students who are paid for grades tend to complete more credits and perform moderately better. But, there is no significant evidence that students increase performance in response to pay for performance on exams (although the empirical methodology is design to identify effects larger than 0.15 standard deviations, and may therefore miss smaller effects which may still be large enough to justify use of such programs).

[7] Although, in a competitive setting it is not clear how much of a subject's behavior is driven by the financial incentives and how much is driven by an inherent desire to perform well relative to others, regardless of the financial benefit of doing so. Cotton, McIntyre, and Price [2013] presents evidence that simply framing a task as a contest can lead to better performance by some participants. In our study, we frame a common task as *two different contests*, which allows us to pick up on differences across alternative allocation mechanisms.

There is a continuum of heterogeneous students of total mass one. Each student belongs to one of two demographic groups, $\mathcal{A}$ (for "advantaged") or $\mathcal{D}$ (for "disadvantaged"), where μ denotes the mass of the disadvantaged group.[8] Each student has a privately-known cost type $\theta \in [\underline{\theta}, \overline{\theta}]$. Within group $j = \mathcal{A}, \mathcal{D}$ the distribution of types follows $\Theta \sim F_j(\theta)$, where $F_{\mathcal{D}}$ stochastically dominates $F_{\mathcal{A}}$ according to the likelihood ratio ordering, and F_j admits a strictly positive density $f_j(\theta),\ j = \mathcal{D}, \mathcal{A}$.[9] For convenience, we denote the unconditional type distribution by $F_{\mathcal{K}}(\theta) \equiv (1-\mu)F_{\mathcal{A}}(\theta) + \mu F_{\mathcal{D}}(\theta)$.

Each student chooses some human capital level from the set $\mathcal{S} \in [\underline{s}, \infty)$, but in order to produce s units of human capital, she must incur a cost given by $C(s;\theta) = \theta s$, which is strictly increasing in both θ and s.[10] There is mass one of heterogeneous prizes (representing college seats) denoted $\mathcal{P} = [\underline{p}, \overline{p}]$, where $p \in \mathcal{P}$ indexes the quality level of the college. These quality levels are distributed according to $P \sim F_{\mathcal{P}}(P)$.

When a student with human capital s is matched with a college of quality level p, a match utility $U(p,s)$ is realized. Allocations of college seats are determined within a frictionless market whose stable equilibrium is characterized by assortative matching on human capital and college quality. The decentralized assortative equilibrium can be represented by a centralized mechanism which uses measured human capital to allocate matching rights to students in rank-order fashion. In this paper we concentrate on two specific mechanisms: color-blind allocations and representative quotas, and we differentiate equilibrium objects tied to these mechanisms by superscripts cb and q, respectively.

Let $G_j^r(s),\ j = \mathcal{A}, \mathcal{D}, \mathcal{K}$ denote the human capital distributions arising in equilibrium from group j under mechanism r, and let $P_j^r(s) = p$ denote the function which determines the college quality level allocated by a mechanism to a student from group j with output level s. A color-blind mechanism matches students and schools assortatively without taking demographics into account. In the continuum model, this implies a mapping between the quantiles of $G_{\mathcal{K}}^{cb}$ and $F_{\mathcal{P}}$ in the following way:

$$P_{\mathcal{A}}^{cb}(s) = P_{\mathcal{D}}^{cb}(s) = P^{cb}(s) = F_{\mathcal{P}}^{-1}\left[G_{\mathcal{K}}^{cb}(s)\right].$$

[8] Bodoh-Creed and Hickman [2014] actually begin by fleshing out a model with finitely many agents. They then show that when the number of agents and college seats is large the analytically unruly equilibrium of this model is well approximated by a simpler setting where agents and college seats each come from a continuum. For expositional simplicity in this paper, we begin with the continuum setting. The interested reader is directed to Bodoh-Creed and Hickman [2014] for further details on the continuum approximation to large, finite markets resembling the one presented here.

[9] Likelihood ratio dominance is a refinement of first-order dominance, with the added condition that $F_{\mathcal{D}}(\theta|\theta \in B)$ dominates $F_{\mathcal{A}}(\theta|\theta \in B)$ in the first-order sense for any measurable subset $B \in [\underline{\theta}, \overline{\theta}]$.

[10] Costs can be thought of as the shadow value of time arising from a labor-leisure trade-off.

In words, the mechanism determines the quantile rank of s within the overall human capital distribution, and then matches a student having s with a college at the corresponding quantile rank. For example, the 75$^{\text{th}}$ percentile student matches with the 75$^{\text{th}}$ percentile college, the median to the median, and so on.

The decentralized market may also involve a preference for demographic diversity (on the part of colleges), such that deviations from assortative matching across groups are possible, but within groups matching remains assortative. For example, a fully representative demographic quota begins by earmarking college seats specifically for the disadvantaged group, and in such a way that all the moments of the two resulting sets of prizes are the same. Thus, it splits the college admissions contest into two separate contests where disadvantaged students compete only among themselves for mass μ of seats following distribution $F_{\mathcal{P}}(p)$, and the advantaged group students compete only among themselves for mass $(1-\mu)$ of seats which also follow distribution $F_{\mathcal{P}}(p)$. Since allocations within each group are still assortative, the allocation rule which implements a quota mechanism is

$$P_j^q(s) = F_{\mathcal{P}}^{-1}\left[G_j^q(s)\right],\ j = \mathcal{A}, \mathcal{D},$$

or in other words, the rule matches *group-specific* quantiles in s with the corresponding quantiles in $\mathcal{P}$. Essentially then, from a student's perspective the distinguishing characteristic of a quota mechanism is that it alters the distribution of one's competitors, while leaving all other aspects of the contest the same as under a color-blind rule.

At the end of the game, the payoff to an agent in group $j = \mathcal{D},\ \mathcal{A}$ with human capital s under mechanism $r = cb, q$ is the match utility minus the cost of achievement, or

$$\Pi_j^r(s;\theta) = U\left(P_j^r(s), s\right) - C(s;\theta).$$

Before investing in human capital, students observe the prize distribution $F_{\mathcal{P}}$, the admission rule, $r \in \{cb, q\}$, and the mass of competitors from each group μ. Under the payoff mapping induced by a particular admission rule $P_{\mathcal{A}}^r(s)$ and $P_{\mathcal{D}}^r(s)$, students optimally choose their achievement level based on their type and the types of potential match partners, taking into account opponents' optimal behavior. The model defined above is an asymmetric, multi-object, all-pay auction with single-unit demands. A set of equilibrium investment functions $s^* = \sigma_j^r(\theta)$ arise which generate optimal human capital choices. These are defined by a first-order condition

$$U_1\left[P_j^r(s), s\right]\frac{\partial P_j^r(s)}{\partial s} + U_2\left[P_j^r(s), s\right] = C'(s;\theta),\ \ j = \mathcal{A}, \mathcal{D},\ r = cb, q.$$

The first order conditions intuitively depict the different components of a student's human capital investment incentives. The second term on the left-hand side, $U_2\left[P_j^r(s), s\right]$, is the direct benefit of having an additional unit of productive human capital. The first term on the right, $U_1\left[P_j^r(s), s\right]\frac{\partial P_j^r(s)}{\partial s}$ is the indirect benefit which comes through competition with other students: if a student produces more human capital (holding fixed the outputs of her competitors), then she will be matched to a better quality school. Of course, the sum of these two benefits are equilibrated in equilibrium with the marginal cost of human capital on the right-hand side.

Bodoh-Creed and Hickman [2014] showed that when $F_{\mathcal{D}}$ dominates $F_{\mathcal{A}}$ according to the likelihood ratio order, then a quota will induce top students in group $\mathcal{D}$ (group $\mathcal{A}$) to decrease (increase) investment, and middle- and low-ability students in group $\mathcal{D}$ (group $\mathcal{A}$) to increase (decrease) it. Attaching a sign to the average change may depend on the parameters of the model but in general, average investment among the disadvantaged group will rise. The effect may go either way for the advantaged group.[11]

3.1. **Numerical Example: Color-Blind versus Quotas.** To illustrate this phenomenon, we present two numerical examples with Cobb-Douglas match utilities $U(p,s) = p^{\alpha_p}s^{\alpha_s}$.[12] In Example 1, we set the Cobb-Douglas utility parameters to $(\alpha_p, \alpha_s) = (0.15, 0.75)$, so that one's own human capital is more valued than the quality of one's match partner, and in Example 2, we set $(\alpha_p, \alpha_s) = (0.75, 0.15)$, so that the reverse is true. The equilibrium investment functions under both mechanisms are depicted in Figure 2. In both examples investment in the disadvantaged group increases under a quota, whereas it rises on average for the advantaged group in Example 1 and falls in Example 2.

Figures 2 and 3 depict the investment patterns predicted by theory. The two figures show a comparison of the investment functions and human capital CDFs for each of the two groups. In each case, there is a single interior crossing point, and the upper bounds of the two distributions are different. For the disadvantaged group, the upper bound of human capital attainment under a quota is smaller than that under a color-blind rule, because a positive mass of the top students reduce investment, while middle- and high-cost students increase it. Intuitively, the policy aids the top students from

[11]More concretely, likelihood ratio dominance implies that the two type densities have a unique interior crossing point, and if the type densities are both unimodal, then the crossing point in the investment functions for the disadvantaged group, $\sigma_{\mathcal{D}}(s)$ and $\sigma_{\mathcal{A}}(s)$ will occur to the left of the density crossing point, which must also be to the left of the mode of $f_{\mathcal{D}}$. This will generally be associated with a mean increase in minority investment under a quota.

[12]In both examples, $[\underline{\theta}, \bar{\theta}] = [1,2]$, $\mu = 0.5$, $F_{\mathcal{P}}$ is $Uniform(1,2)$, and the type distributions are truncated normals with standard deviation 0.15, with $F_{\mathcal{D}}$ having mean 1.5 and $F_{\mathcal{A}}$ having mean 1.1 (so that the two distributions are ordered by likelihood ratio dominance, see Figure 1).

FIGURE 1. NUMERICAL EXAMPLES: TYPE DENSITIES

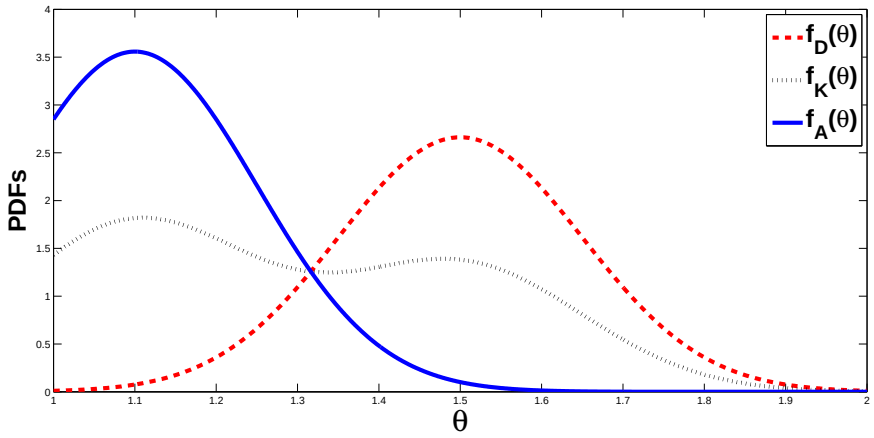
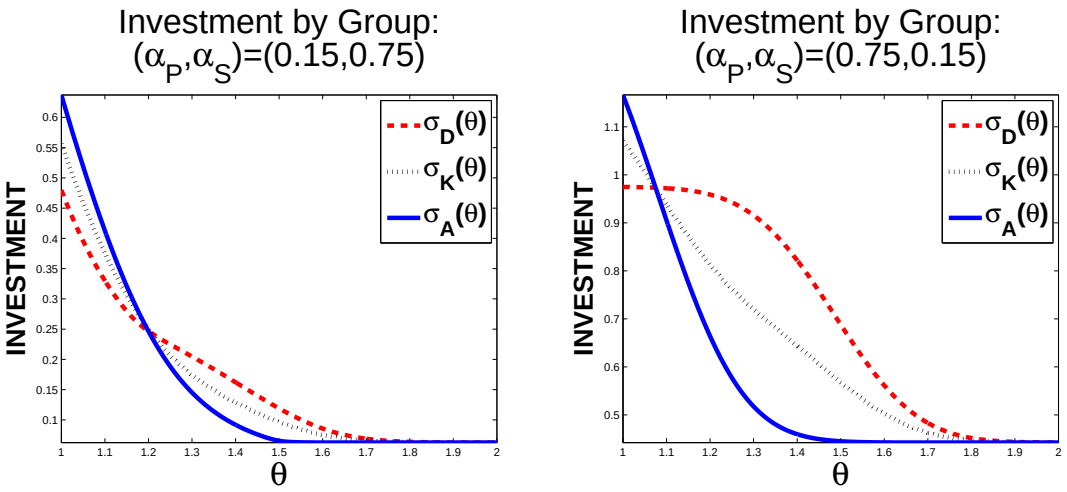


FIGURE 2. INVESTMENT STRATEGIES: COLOR-BLIND VS QUOTA

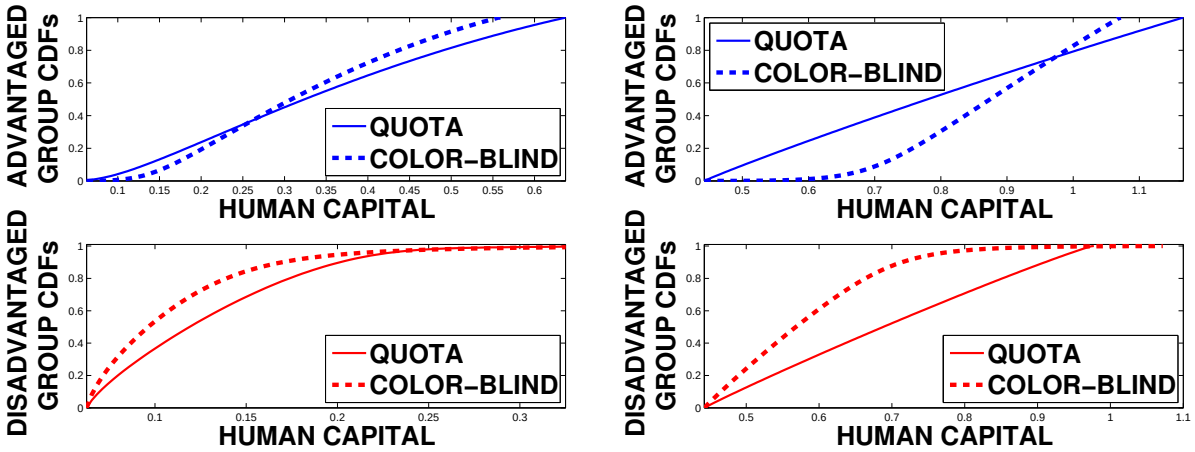


(A) Example 1 (B) Example 2

group $\mathcal{D}$, but since their outcomes were already close to the upper bound for a color-blind rule, they rationally reduce effort. For other students in group $\mathcal{D}$, the policy alleviates so-called "discouragement effects" by making them competitive for higher quality outcomes, and they respond by increasing investment.[13] The opposite is true for investment in the advantaged group. In both examples, AA clearly increases average human capital investment by the disadvantaged group. In the experimental analysis, we show that AA leads to changes in the performance distribution consistent with these

[13]Discouragement effects are a common phenomenon in contests and all-pay settings. The term refers to the idea that, holding one's own type fixed, if the distribution of competition shifts so that one slips further behind, then one will eventually decrease effort.

FIGURE 3. HUMAN CAPITAL DISTRIBUTIONS: COLOR-BLIND VS QUOTA



(A) Example 1

(B) Example 2

theoretical predictions, and we also investigate its effects on intermediate inputs which lead to higher levels of measured human capital as well.

4. EXPERIMENTAL DESIGN

We designed our experiment so that our incentives would be based on studentsâĂŹ performance on the American Mathematics Competition 8 test (AMC8), a national test conducted every year in November for students in 8^{th} grade and below. Participation in the AMC8 is decided on the school level. The AMC8 consists of 25 questions taken in a 40 minute period. We conducted our experiment during the 2012 and 2013 AMC8 competitions, for which the median national score was 10 / 25 and the 99th percentile student scored 22 / 25. The AMC8 website explains: "The examination provides an opportunity to apply the concepts taught at the junior high level to problems which not only range from easy to difficult but also cover a wide range of applications. Many problems are designed to challenge students and to offer problem solving experiences beyond those provided in most junior high school mathematics classes."

4.1. **Partner Schools.** Our total sample includes 992 students in 5th through 8th grade from 10 different schools in Utah County, Utah, including both charter schools and regular public schools. All students from participating classrooms in our partner schools were involved in the study on an opt-out basis, which provided us with a sample of test subjects who were broadly similar to the demographic mix at their school. On some dimensions, socioeconomic characteristics for our schools are comparable to the nation as a whole, though not on all. We estimate a median household income of $59,800,

compared to a nationwide median of $53,046.[14] In 2012, approximately 33 percent of students in our sample were eligible for free or reduced-price lunch, compared to a national average of 48 percent and a statewide average of 38 percent. That year the elementary schools in our sample performed significantly better than other Utah schools in terms of meeting state math standards (approximately 91% vs 76%), while the middle schools in our sample performed slightly worse (81% vs 83%). Most of our schools had participated in the AMC8 in the past before partnering with us for this study.

Our partner schools exhibited a high degree of racial and cultural homogeneity. Less than 1 percent of students were black and only 7.5 percent were Hispanic, compared to nationwide averages of 15.25 percent and 22.2 percent for blacks and Hispanics, respectively. All schools in this study serve suburban populations. However, given that the goal of the current exercise is to cleanly test theory of incentives, the demographic homogeneity in our sample may actually be an advantage. We chose grade-level as our characteristic on which to base the affirmative action treatment, which ensures that the groups of advantaged and disadvantaged students only differ in observable ways (i.e., age and grade) but are otherwise similar. Focusing on such racially homogeneous districts allows us to largely rule out other cultural/behavioral phenomena (such as stereotype threat) which might confound the pure incentive effects arising from our exogenous policy variation.

4.2. **Treatment Groups and Incentives.** Participants in our study first took a practice exam using an AMC8 test from a previous year. We used this practice test as a baseline measure of each student's ability. We randomized each individual student into one of two treatment groups that were analogous to "color-blind" and AA (quota) policies. We ran competitions involving both 5^{th} and 6^{th} graders, and separate competitions involving 7^{th} and 8^{th} graders. Students in the lower grades (5 and 7) are henceforth referred to as the "disadvantaged" demographic, and students in the higher grades (6 and 8) are referred to as the "advantaged" demographic group, given that they are one year older and have received one more year of mathematics education on average.[15] Those in the quota treatment were assigned to compete only against other students in their own grade

[14]These figures are based on information downloaded from http://nces.ed.gov/surveys/sdds/framework/tables.aspx

[15]The difference in average preparation between our disadvantaged and advantaged groups is likely to be lower than national differences between black and white students in the US. Using performance figures from the National Assessment of Educational Progress (NAEP) exam, we estimate that by 8th grade, the average black student performs roughly as well on the mathematical portion of the NAEP exam as the average white student would have performed in 6th; this represents a performance gap consistent with two years education difference. The NAEP exam is only done with 4th and 8th grade students, so our estimate is based on a calculation of annual change in ability based on the four year trend.

(5^{th}, 6^{th}, 7^{th} and 8^{th} grade separately). This is equivalent to taking half of the prizes from a race-neutral treatment, and reserving them for disadvantaged students.

Each student received a sheet of paper that described the group that they were assigned to, how many students they would be competing against, the score distribution of the students in their group based on the practice test, and the prize structure. Students received their practice score back at the same time so that they could easily compare where they fit within the distribution for their competition group. The sheet also described the prize structure. The top 30 percent of students within each competition group received cash prizes, which were uniformly distributed between \$4 and \$34 in \$2 increments with the highest payments going to those with the highest final exam scores. In the color-blind treatments, students competed against others in their own grade and in an adjacent grade. For example, 7^{th} grade students in a color-blind treatment needed to score within the top 30 percent of all 7^{th} and 8^{th} grade students in their treatment to receive a prize.

In the quota treatment, students competed against others in their own grade only. For example, 7^{th} grade students in the quota contest only had to score within the top 30 percent of 7^{th} graders in their treatment to receive a prize. The prize distribution was identical across all competition groups, with each one competing for the same aggregate set of prizes on a per capita basis. Thus, for an advantaged or disadvantaged student of a given ability level, the only difference across the two treatments is the distribution of one's competitors.

We printed information relative to each competition group on a different color of paper so that students could visually see in their classroom that roughly half of the students were assigned to each treatment. Altogether, there were six different groups: four groups for the quota treatment (one for each grade) and two groups for the neutral treatment (one for elementary schools and one for middle schools). In a web appendix, we provide an example of the information sheet that we provided to each group.[16]

4.3. **Math Learning Website.** At the bottom of the information sheet was the url of a website we set up with practice problems drawn from five past AMC8 exams. At 25 questions each, this made for 125 total practice problems covering six different math subjects: Arithmetic, Algebra, Combinatorics, Geometry, Logic, and Probability. Problems were divided into a set of 31 total quizzes. Each year, the 25 AMC8 exam questions are numbered in increasing order of difficulty. For each of the previous five year's exams, the website included one quiz covering problems 1-10, a second quiz covering

[16]Copies of the information sheets given to test subjects are available for download at http://home.uchicago.edu/~hickmanbr/uploads/CHP2014_WEB_APPENDIX.zip

problems 11-20, and a third covering problems 21-25. Test subjects were notified that each grouping of 3 same-year quizzes were ordered by their difficulty level. We also arranged this same battery of math problems into an additional set of 16 quizzes, each containing 5 subject-specific math problems. These subject quizzes were also ordered by their difficulty level.

Students could attempt each quiz as many times as they like, or move on to additional materials they had not yet tried. After completing each quiz, our software displayed an instructional page which reported to each student her score, the correct answers for each problem, and step-by-step solutions published by the developers of the AMC8. Students were provided with a web page that contained links for all of the quizzes we offered, but in order to access the quizzes, they had to input their name, grade, and school on the first page of the web form. This allowed us to track online activities, including which students visited the website, how many different subjects they tried, how much time they spent, how many questions they attempted, what they answered on each attempt at each question, and how much time they spent viewing the instructional page.

Within each quiz, questions were separated on different web pages in blocks of 3, 4, or 5 questions per page, and the instructional page at the end displayed feedback for all questions on a single page.[17] Time on our website was measured at the page level, meaning that we got a time measure for blocks of either 3, 4, or 5 questions. In order to convert this information into a time spent per question measure, we divided each block-level time observation by the number of questions within that block. Instructional page times for 10-question quizzes were split into two observations a piece by dividing by two in order to make them comparable to 5-question instructional page view times.

One difficulty arose in that there were clear instances where students left the website in the middle of a quiz for several hours or more. To adjust for this problem we chose truncation points on the domains of time per question and instructional page view time, and we replaced each observation above that point with the appropriate student-specific censored mean.[18] In selecting our truncation point we looked for occurrences of "holes" in the support of the distribution of times per question.[19] For our time per question data,

[17]Each 10-question quiz was broken into three pages with 3 questions on page one, 3 questions on page two and 4 questions on page three. Each 5-question quiz displayed all 5 problems on a single page.

[18]To illustrate this rule, suppose that Tommy attempted three 5-question quizzes for a total of 15 questions. Suppose further that we observed times of 5 minutes each for seven questions, 15 minutes each for another seven questions, and 2000 minutes for the last one. Then if the truncation point were, say 30 minutes, the fourteenth observation of 2000 is replaced by Tommy's idiosyncratic censored mean time of 10 minutes (for all other questions he attempted). As a robustness check, we also ran our analysis by simply dropping truncated observations in stead, and results are very similar to those we present below.

[19]More specifically, a hole in the distribution support was defined as the lowest point at which a full-support condition fails, which we estimated as the lowest point where a kernel-smoothed density estimate hit zero. The idea behind this rule is that if the type distribution has full support, then the distribution

this leads to a truncation point of 26.14 minutes per question (the 99.35^{th} percentile), and for instructional page views, we get a truncation point of 108.39 minutes (the 98^{th} percentile). In the Appendix we display a histogram of time spent per question and instructional page view times, including observations above and below the truncation point.

At the end of the day, the time monitoring capability on our website is not perfect, and it is impossible to directly observe work stoppages in the middle of a quiz question. In particular, it may still be the case that smaller work stoppages occur below the truncation points. Therefore, in terms of time per question we are effectively interpreting work stoppages of less than 27 minutes as time which comes at a positive cost to the child. We argue that 27 minutes is a reasonable truncation point for several reasons. First, work stoppages for our uncensored time observations (most of which were less than 10 minutes) would serve as a poor substitute for longer, unbroken leisure spells. Second, since this potential problem is the same across both treatment groups, there is no reason to believe that our results are being aided by it.

Third, the AMC8 contains fairly challenging material that may require significant time inputs for some students. Table 6 in the Appendix displays the mean and variance of time spent per question attempt, using the censored sample of times. The most difficult subject appears to be combinatorics, with a mean time of 2.839 minutes and a standard deviation of 3.532. Given that the censored distribution of time per question is right skewed, and 10 minutes (the 98^{th} percentile of the un-censored sample) is roughly two standard deviations above the mean for combinatorics, it is plausible that roughly 1.5% of our sample could exist on the interval between 10 minutes and 26 minutes.

As for instructional page view times, it is informative to consider a particular student whom we will rename "Kate" to protect her identity. Kate, a 7^{th} grader in the quota treatment, spent more time than anyone else on the website (after time adjustments), averaging roughly 53 minutes per day during the study period. At 55 question attempts, Kate was also above the 94^{th} percentile on that dimension as well. She attempted 11 quizzes of 5 questions each, averaging 2.24 minutes per question attempt and 7.49 minutes per question on the instructional page with solutions. She spent an hour or more on 5 of her quiz attempts, each time spending the majority of her time on the instructional page. For each of Kate's 11 quiz attempts we see that she clicked through to the quiz termination screen herself. Kate is an example of a student who displayed consistent patterns of substantial time inputs into many of the quizzes she took—particularly on the instructional page—while never having left a quiz session open overnight. See the

of times per question should have full support as well since the choice of how much time to spend is continuous. For a more complete description of our truncation point selection rule, see the Appendix.

appendix for further explanation and summary statistics concerning our time truncation rules.

4.4. **Testing.** Students took the actual AMC8 test in their regular classrooms, under all of the normal conditions in which students around the country take the AMC8. Most of the students in our study attended schools where participation in the AMC8 was already being offered to students by their teachers, but on a voluntary basis. The schools that cooperated with us for our study administered the test to all students within each participating classroom on an opt-out basis, so that all students participated in the study, except those whose parents proactively signed and returned an opt out form. The study involved two in-class exam sessions: the practice test was the AMC8 exam for the previous year, and the final exam was the AMC8 for the current year. We graded each student's final exam ourselves, before their answer sheets were submitted to the AMC office for official processing. We then assigned prizes to each student based on her score and treatment group. The cash prizes where delivered to each school shortly after the final exam, and handed out to each student in an envelope. The outcome measures that we use in the next section include both the effort-based measures with website data, as well as a performance-based measure based on the students' scores on the AMC8.

5. Results and Discussion

5.1. **Descriptive Statistics.** In this section we present the empirical results of our field experiment. Table 1 contains descriptive statistics on our test subjects. Roughly three quarters of our sample were $7^{th}/8^{th}$ graders. The difference between these and our $5^{th}/6^{th}$ grade test subjects is that the latter all came from accelerated classes, whereas the former are representative of the overall student body within their respective schools.[20] This difference is born out in the data: while 8^{th} grade students did best on the pre-test with an average score of 9.04, 6^{th} graders as a group came in second at 8.12 on average. 7^{th} and 5^{th} grade average pre-test scores are close, at 7.58 and 7.19, respectively.[21]

We have also broken down test scores by two groups that we refer to as *investors*—students who logged on to our website at least once during the study period—and *non-investors*—those who did not. Students who did better on the pre-test were more likely

[20]Since the AMC8 is a challenging exam targeted toward students up to 8^{th} grade, our partner schools only offer it to accelerated students in $5^{th}/6^{th}$ grade.

[21]The national AMC8 population in 2013 (statistics downloaded from https://amc-reg.maa.org/reports/generalreports.aspx) had mean and median of 10.69 and 10 out of 25, with standard deviation of 4.44. These figures are illustrative of the difficulty of the exam, but the comparison to our sample population is not altogether straightforward though. The AMC8 is predominantly administered on an opt-in basis, whereas our experiment was administered on an opt-out basis (meaning all students in participating classrooms were involved unless they requested not to be in writing).

to invest their time into learning math, although some students who did not do as well also chose to invest, and many students who did quite well on the pre-test choose not to invest. For the group of investors, we also present summary statistics concerning their activities on the website. Investors' times ranged between a few minutes and 8.92 hours, or an average of about 53 minutes per day over the study period. Number of questions attempted ranged between 1 and 120, with mean and standard deviation of roughly 19 and 23, respectively. Subjects represents the number of different subject categories a student attempted, using the subject-specific quizzes, being about two on average.

5.2. **Empirical Analysis.**

5.2.1. *Testing Overall Differences by Treatment.* Tables 2 and 3 investigate the effect of a quota on the overall population, including both advantaged *and* disadvantaged groups. The first column of Table 2 displays the mean of a binary variable, being the fraction of test subjects from each treatment group who logged on to our website at least once to practice math. The results indicate that students in the quota treatment were 75% more likely to have visited the website than students in the color-blind treatment. They also tried out more subjects, spent more time on the website and answered more questions. As for the investment variables, the reader should keep in mind that Tables 2 – 4 aim to measure a treatment effect of a policy on both the intensive and extensive margins of investment. This is why the effort numbers in Table 2 and afterward appear small: *they are averaged over both investors and non-investors*. Table 2 indicates that students in both treatments scored roughly the same on the final exam. This is allowed for by the theory, where predictions for the overall population are qualitatively ambiguous, but later on we will see a different story when we condition on demographic group.

Table 3 provides statistical tests for the raw differences displayed in table 2. In the first row we run a simple regression using a dummy for the quota treatment, meaning it represents the experimental difference between a quota rule and color-blind allocations. Each cell in the table represents a separate regression with the outcome variable labeled in the column header. We report the point estimate and p-value for a test of the hypothesis that the coefficient equals zero (*i.e.*, that the quota treatment makes no difference). From the table we see strong evidence that AA increases the fraction of students willing to invest at least some time. We also see evidence that it induces them to experiment with more subjects, as well as increase the total time invested and number of questions attempted. Although these last two differences are only marginally significant, the estimated magnitudes are large, with quota students logging an estimated 57% and 70% more inputs of time and question attempts, respectively.

TABLE 1. STUDENT DESCRIPTIVE STATISTICS

	Mean	Median	Std. Dev.	N
Pre-Exam Scores				
All	8.45	8	2.90	992
5th Grade	7.19	7	2.39	48
6th Grade	8.12	8	2.47	155
7th Grade	7.58	7	2.84	275
8th Grade	9.04	9	2.82	396
Investors	9.46	10	3.19	118
Non-Investors	8.32	8	2.83	874
Final Exam Scores				
All	8.64	8	2.88	895
5th Grade	7.40	7	2.22	42
6th Grade	9.17	9	2.82	133
7th Grade	8.12	8	2.90	233
8th Grade	8.75	9	2.80	374
Investors	9.20	9	3.06	113
Non-Investors	8.56	8	2.84	782
Human Capital Investment (Investors Only)				
Total Time	43.65	26.85	64.65	118
Problem Solving Time	32.99	19.31	41.43	118
Instructional Time	10.66	3.37	38.85	118
Questions	18.89	10.00	22.53	118
Subjects	1.94	1.00	1.43	118

Notes: National AMC8 Statistics for 2013 were downloaded from https://amc-reg.maa.org/reports/generalreports.aspx. All time figures are post-censoring as described in Section 4.3 and quoted in minute units.

TABLE 2. EFFORT AND PERFORMANCE BY TREATMENT

	Investment				**Performance**
	Used Website	# Subjects Attempted	Total Time	# Questions Attempted	Final Exam Score
Quota	0.154	0.284	6.634	2.729	8.680
Std. Err.	(0.015)	(0.037)	(1.216)	(0.456)	(0.139)
Neutral	0.088	0.189	3.932	1.817	8.604
Std. Err.	(0.014)	(0.035)	(1.149)	(0.431)	(0.133)
N	992	992	992	992	895

Notes: Each cell provides the mean of the measure listed in each column. Standard errors are provided in brackets. Estimates under each of the four effort variables are intended to capture the effect of a treatment on human capital investment for the total study population, and are therefore averaged over both investors and non-investors.

In the second and third rows of Table 3 we include additional controls (pre-test score and/or school fixed effects) as a way of checking whether our student-level randomization did what it was supposed to. In general this appears to be true, as they do not create a significant shift in point estimates. Just in case though, we also include these controls in the tables that follow as well.

5.2.2. *Testing Differences by Treatment Within Demographic Groups.* Recall that the theory allows for AA to have differential effects by ability and demographic group, both in signs and magnitudes. In Table 4, we add a demographic dummy to investigate this claim. Each column presents estimates for a regression equation of the form

$$Outcome = \beta_0 + \beta_1 Quota + \beta_2 Advantaged * Quota + \beta_3 Advantaged + \beta_4 Ability + U,$$

where *Quota* is a dummy for treatment status, *Advantaged* is a demographic dummy, *Ability* is a student's standardized pre-test score, and the specific *Outcome* variable is labeled in the column header. With the inclusion of the interaction term $Advantaged * Quota$, the coefficient β_1 represents the average effect of AA specifically on the disadvantaged group. The effect of the policy on the advantaged group is represented by the sum $\beta_1 + \beta_2$. For completeness, all regressions include controls for school-level fixed effects, and for the primary effects of interest we report p-values in brackets.

TABLE 3. TESTING DIFFERENCES BY TREATMENT

	Investment				**Performance**
	Used Website	# Subjects Attempted	Total Time	# Questions Attempted	Final Exam Score
Quota − Neutral	0.066***	0.095*	2.701	0.912	0.076
P-Value:	*[0.001]*	*[0.061]*	*[0.107]*	*[0.146]*	*[0.693]*
(*Controls: none*)					
Quota − Neutral	0.065***	0.093*	2.650	0.884	0.097
P-Value:	*[0.002]*	*[0.067]*	*[0.113]*	*[0.158]*	*[0.576]*
(*Controls: pre-test scores*)					
Quota − Neutral	0.058***	0.078	2.404	0.773	0.164
P-Value:	*[0.005]*	*[0.130]*	*[0.158]*	*[0.224]*	*[0.346]*
(*Controls: pre-test scores, school FEs*)					
N	992	992	992	992	895

Notes: Each cell represents a separate regression. The number reported is the coefficient for the quota treatment. Row 1 includes no controls and provides a statistical test of the differences in Table 1. Row 2 includes control for practice test score. Row 3 includes school fixed effects. P-values for a two-sided test of the null hypothesis of zero difference are italicized and in brackets. Estimates under each of the four effort variables are intended to capture the effect of a treatment on human capital investment for the total study population, and are therefore averaged over both investors and non-investors.

For disadvantaged group students we find evidence of large and positive effects across all four investment measures. First, we see a highly significant 8.7 percentage point increase in disadvantaged students' willingness to spend at least some time on the website. To put this in perspective, we can compute a within-demographic percent change for the disadvantaged group by $100 * (\beta_1/\beta_0)\%$, which amounts to an increase of 119% on the extensive margin, relative to their disadvantaged counterparts under the color-blind treatment. We also see a significant and even larger increase in terms of time investment: disadvantaged students under the quota treatment increased investment by 181%. The other two measures capture specific tasks done during time spent on the

website: # of subjects attempted and # of questions attempted. Although the latter is only marginally significant, both render large point estimates for increases of 101% and 100%, respectively.

Another striking feature of the table is the performance measure. We find a large and significant difference on final exam scores for disadvantaged students in the quota treatment: they are estimated to have lifted their scores by 0.624 AMC8 points, or a remarkable 21.7% of a standard deviation over their disadvantaged counterparts in the control group. Although some portion of this effect may also be due to increased effort and concentration on the day of the final exam, we interpret this and the other results in Table 4 as evidence that AA can be an effective tool for providing disadvantaged students with increased incentives to acquire pre-college human capital through investment.

However, one concern is that strengthening incentives for one demographic group may come by weakening them for the other. Table 4 provides evidence that these gains for disadvantaged students come at little or no cost in terms of average human capital investment among advantaged students. All point estimates on β_2 are negative, but for 4 out of 5 outcome measures it is smaller in magnitude than β_1. For the final exam outcome, the sum of the two coefficients is slightly negative (representing about 3% of a standard deviation) but with a large p-value. The outcome measure under which $\beta_1 + \beta_2$ is most significant is the binary measure of investment, with a p-value of 0.148. This implies an estimated percent change of $100 * (\beta_1 + \beta_2)/(\beta_0 + \beta_3) = 36.9\%$ on the extensive margin for advantaged students under a quota. Thus, we do not find evidence that there is a trade-off between average human capital investment across demographic groups; if anything the data seem to slightly favor a small increase of investment for the advantaged demographic as well. Figures 8 – 11 in the appendix contain graphical depictions of the distributional shifts of inputs and outputs by demographic and treatment group.

5.2.3. *Selective Attrition.* One potential source of bias in our results concerning the performance measure (final exam score) is that 97 of the students who took the practice test and were randomly assigned to a competition group ended up not taking the final test.[22] We find that among the disadvantaged students, those assigned to the quota group were less likely to miss the final exam (10.6% vs. 16.7%). We also find that among the students who didn't show up for the final test, the disadvantaged students assigned to the quota group had higher practice scores than the disadvantaged students not assigned to the quota group (7.16 vs 6.35). However, the practice scores among the students who did show up for the final test were nearly the same across these two groups (7.91 vs 7.79).

[22]Note that this problem does not arise with the four investment measures, which did not require observing a final score for us to observe. Note that in Table 4, the sample size for the first four columns represent the full sample of test subjects.

TABLE 4. TESTING DIFFERENCES BY DEMOGRAPHICS AND TREATMENT

	Investment				**Performance**
	Used Website	# Subjects Attempted	Total Time	# Questions Attempted	Final Exam Score
Constant ($\widehat{\beta}_0$)	0.073***	0.144**	3.033	1.316*	8.147***
Std. Err.	(0.024)	(0.059)	(1.982)	(0.741)	(0.209)
Quota ($\widehat{\beta}_1$)	0.087***	0.146*	5.517**	1.312	0.624**
Std. Err.	(0.033)	(0.083)	(2.757)	(1.030)	(0.287)
P-Value:	*[0.009]*	*[0.077]*	*[0.046]*	*[0.203]*	*[0.030]*
Advantaged $*$ *Quota* ($\widehat{\beta}_2$)	-0.047	-0.111	-5.034	-0.866	-0.712**
Std. Err.	(0.042)	(0.105)	(3.506)	(1.310)	(0.360)
Advantaged ($\widehat{\beta}_3$)	0.028	0.083	1.613	0.877	0.488*
Std. Err.	(0.030)	(0.076)	(2.545)	(0.951)	(0.264)
Ability ($\widehat{\beta}_4$)	0.029***	0.045*	1.009	0.572*	1.280***
Std. Err.	(0.011)	(0.027)	(0.893)	(0.334)	(0.092)
School Fixed Effects	yes	yes	yes	yes	yes
N	992	992	992	992	895
Additional Test: Effect of Quota on Advantaged Group					
$\widehat{\beta}_1 + \widehat{\beta}_2$	0.040	0.035	0.483	0.446	-0.089
P-Value:	*[0.123]*	*[0.586]*	*[0.823]*	*[0.581]*	*[0.684]*

Notes: Each column is a separate regression. Advantaged is an indicator variable for whether the student is a 6^{th} or 8^{th} grader (the older group in each school type). Ability is the standardized pre-test score, where standardization is based on the mean and variance within each school type (*i.e.*, $5^{th}/6^{th}$ or $7^{th}/8^{th}$). Standard errors are in parentheses; p-values for a two-sided test of the null hypothesis of zero effect are italicized and in brackets. Estimates under each of the four effort variables are intended to capture the effect of a treatment on human capital investment for the total study population, and are therefore averaged over both investors and non-investors.

These comparisons all point in a direction opposite of our main results and suggest that the effect of the quota on final performance for disadvantaged students may have been greater in the absence of this selective attrition.

5.2.4. *Policy Responses by Ability Level.* In this section we turn to a more direct investigation of the predictions of theory concerning distributions of test scores that should arise in equilibrium under each allocation mechanism. Specifically, the theory model of Bodoh-Creed and Hickman [2014] predicts that if the underlying cost types for the disadvantaged group stochastically dominate those in the advantaged group, then qualitative patterns like those displayed in Figure 3 should appear. Namely, for the disadvantaged group the test score distributions under a quota and color-blind mechanism should have an interior crossing point, with the former strictly above the latter to the right of the crossing point, and strictly below to the left. In other words, there should be a positive mass of the best disadvantaged group students who decrease output, while students of medium and low ability from that group increase output. The theory makes the opposite predictions for the advantaged group.[23]

While it is impossible to directly observe the distributions of cost types, we can take queues from the distribution of pre-test scores by demographic group, since they reflect how much progress each student will need to increase her payout. We can then examine the distributions of final exam scores within demographic groups under different treatments to see whether our experimental data seem to be consistent with the theory of incentive effects under AA. Figures 4 – 6 depict these comparisons in three plots of empirical cumulative distribution functions for pre-test and final exam scores for grades 7 and 8.[24] For the sake of comparability, we have limited our sample in these figures to include only students for whom we have both test scores. Therefore, Figure 4 plots empirical pre-test CDFs only for 7^{th} and 8^{th} graders who took the final exam.

Figure 4 strongly supports stochastic dominance of initial math proficiency levels across demographic groups.[25] A two-sample Kolmogorov-Smirnov (KS) test rejects the null hypothesis that the disadvantaged and advantaged group distributions are the same, instead favoring of a one-sided alternative hypothesis that the latter stochastically dominates (in the first order sense) with a p-value of 1.03×10^{-5}. This means that the

[23]Although the figures we present in this paper explore distributions of both inputs and outputs, it is important to remember that the predictions of the theory only directly apply to exam score, as this is the variable on which prize allocations are based. The mapping from inputs to outputs may vary by student if each one differs by raw math talent and leisure preference.

[24]We excluded grades 5 and 6 from this exercise since our 5^{th} grade sample was small.

[25]Note that the theoretical predictions we test here actually require likelihood ratio dominance (a stronger form of first-order stochastic dominance) of the cost type distributions; however, since we have only an indirect measure of costs to work with here, we simply test for first-order dominance of the pre-test distributions.

disadvantaged group on average had to achieve more progress in order to be competitive for a prize. Once again, this is not exactly the same as observing costs, but the two are certainly related and the idea of stochastic dominance in cost types appears plausible.

We find evidence in Figure 5 that by the end of the study period the score distribution within the disadvantaged group had diverged by treatment status. A two-sided KS test for disadvantaged group final exams across treatments results in a p-value of 0.105, providing marginally significant evidence that the two distributions were not the same.[26] There appears to be a hint of divergence by treatment group among 8^{th} graders in Figure 6 as well, though we lack sufficient power for a KS test to distinguish the two distributions apart (the p-value for a two-sided test is 0.889).[27]

In interpreting the figures, one caveat should be kept in mind: the KS test can only indicate that two distributions are not the same, but it does not provide for a test of the specific ordering of two distributions with an interior crossing point on different subsets of the support. Therefore, the qualitative patterns displayed in Figures 5 and 6 are suggestive in nature. However, a striking feature of the plots is the remarkable degree to which they conform to the qualitative patterns predicted by theory and displayed in the numerical examples from Section 3. Both sets of CDFs have a single interior crossing point, with the upper support bound for disadvantaged students under a quota differing by four AMC8 points relative to that under color-blind allocations (15 vs 19). Below the crossing point, we see a substantial increase in measured 7^{th} grade math proficiency under a quota. These patterns are reversed for 8^{th} graders (with maxima of 17 vs 16, respectively), though the effects appear much smaller in magnitude.

5.2.5. *Narrowing Achievement Gaps.* We now conclude analysis of our experimental data with a look at the tendency for AA to narrow achievement gaps across demographic groups. Table 5 displays summary statistics on standardized test scores for the pre-test and final exam, for grades 7 and 8. In the top panel of the table scores were standardized within each exam by subtracting the mean and dividing by the standard deviation for

[26]Figure 10 in the appendix contains an additional plot comparing pre-test scores by treatment within the disadvantaged group. The differences in the pre-test distributions are due to selective attrition after we omitted students for whom we have no final exam. The figure suggests that in general selective attrition is working against our result here: the pre-test color-blind distribution is below the pre-test quota distribution for values at or below the median, and and the upper bound of the pre-test distribution for 7^{th} grade quota students is highest. Both characteristics are substantially reversed by the final exam.

[27]Figure 11 in the appendix contains an additional plot comparing pre-test scores by treatment within the advantaged group. The differences in the pre-test distributions are due to selective attrition after we omitted students for whom we have no final exam. Once again, the figure shows that selective attrition is working against finding evidence for the theoretical prediction: the pre-test distributions have a crossing point close to that for the final exam distributions, and their ordering above and below the crossing is reversed, relative to the final exam.

FIGURE 4. PRE-TEST SCORES: 7^{th} GRADE VS 8^{th} GRADE

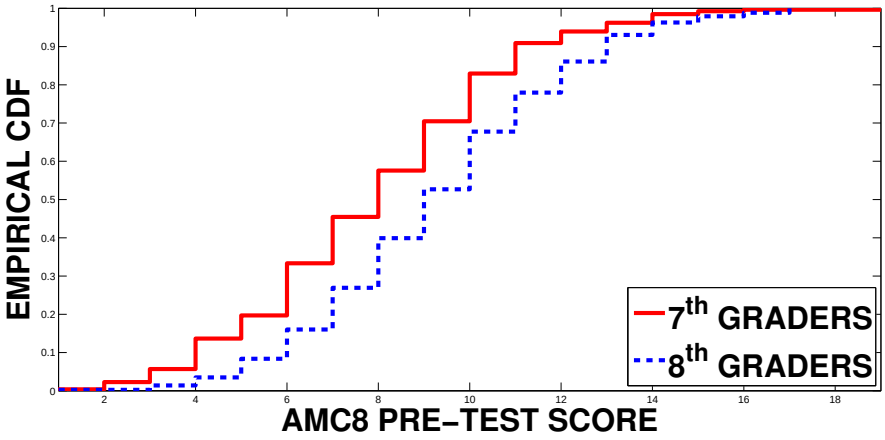


FIGURE 5. SEVENTH GRADE FINAL EXAM SCORES

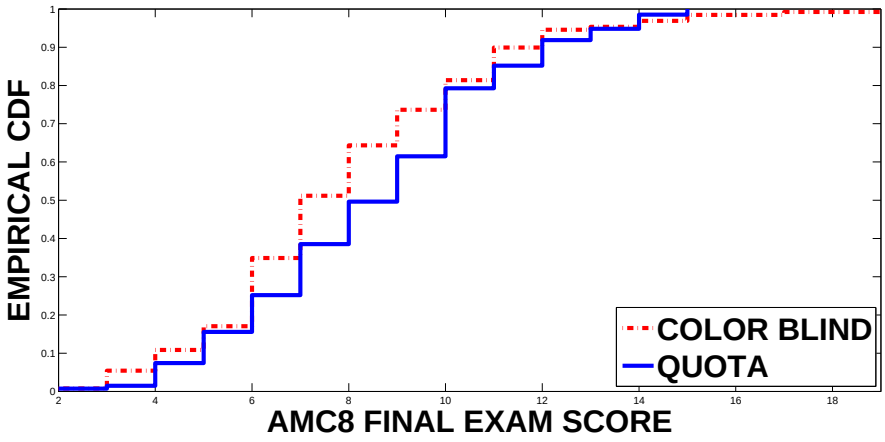
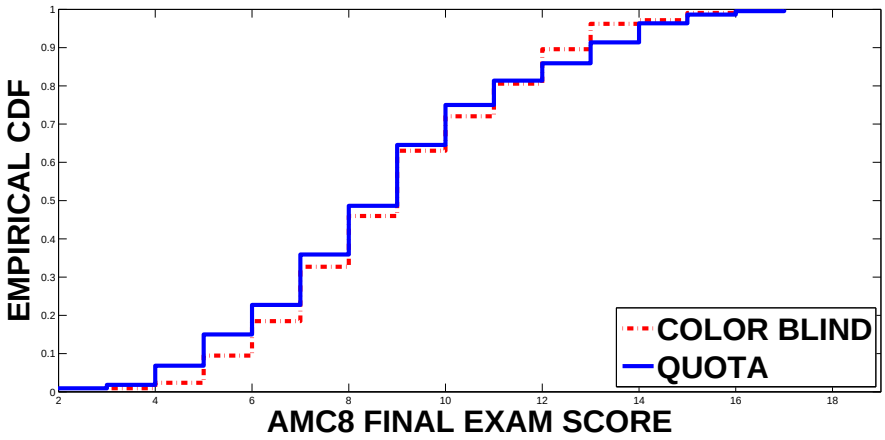


FIGURE 6. EIGHTH GRADE FINAL EXAM SCORES



all grade 7 and 8 students.[28] Therefore, the means indicate how far ahead or behind the population average each demographic group was. We see that 7^{th} grade students were roughly half of a standard deviation behind their 8^{th} grade counterparts on average (or $-0.295 - 0.181 = -0.476$ standard deviations to be exact). However, by the final exam, the gap between 7^{th} and 8^{th} graders had narrowed by about a quarter of a standard deviation (or $-0.138 - 0.085 = -0.223$ standard deviations).

In the lower two panels of Table 5 we break out this effect by treatment group, displaying the same numbers within treatments, but where score standardization now happens within each exam-treatment cell. At least part of the test score convergence had to do with differences between the pre-test and final exam: within the color-blind treatment, about a quarter of the gap (beginning at 0.502) disappeared on the final exam but remained relatively high at 0.367 standard deviations on the final exam. However, the achievement gap under the quota treatment closed substantially more, by about 80%, beginning at 0.442 standard deviations, and ending at only 0.085 standard deviations on the final exam.

The medians tell a slightly stronger story, with the median gap beginning about the same within both treatments, closing virtually to zero under a quota, and closing only slightly under a color-blind mechanism. Finally, observing that the within-treatment-demographic standard deviations are all close to one suggests that the narrowing of gaps within the two treatments was due predominantly to mean/median shifts in test scores. We interpret these findings as evidence that AA can actually help to narrow demographic achievement gaps while equalizing market allocations.

6. Conclusion

We designed a field experiment in which 5^{th} through 8^{th} grade students compete for heterogeneous cash prizes and are paid by their relative performance on a nationwide mathematics exam. Our experimental design creates a microcosm of pre-college human capital investment followed by the college admissions market; namely, voluntary labor-leisure decisions, mathematics learning, and affirmative action. Within this context, the prize allocation rule approximates either a "color-blind" system under which exogenously disadvantaged students compete head to head with more experienced students, or a "quota" affirmative action policy under which a set of prizes are reserved for those in the disadvantaged group. We track student study effort during a 10-day investment period prior to the final exam. Our experimental results support the theoretical predictions of Bodoh-Creed and Hickman [2014], where students strategically adjust their

[28]Once again, in order to make the pre-test and final exam figures comparable, we excluded from the analysis any students whose final scores were missing due to attrition.

TABLE 5. NARROWING GAPS

	Mean	Median	Std. Dev.	N
Achievement Gaps for All Treatments				
Standardized Pre-Score (GRADE 7)	-0.295	-0.267	0.996	264
Standardized Pre-Score (GRADE 8)	0.181	0.071	0.960	431
Standardized Final Score (GRADE 7)	-0.138	-0.205	1.005	264
Standardized Final Score (GRADE 8)	0.085	0.142	0.987	431
Achievement Gaps for Quota Treatment				
Standardized Pre-Score (GRADE 7)	-0.211	-0.199	1.061	135
Standardized Pre-Score (GRADE 8)	0.231	0.135	0.920	220
Standardized Final Score (GRADE 7)	-0.030	0.142	0.961	135
Standardized Final Score (GRADE 8)	0.055	0.142	1.050	220
Achievement Gaps for Color-Blind Treatment				
Standardized Pre-Score (GRADE 7)	-0.243	-0.199	0.904	129
Standardized Pre-Score (GRADE 8)	0.259	0.135	0.982	211
Standardized Final Score (GRADE 7)	-0.251	-0.552	1.041	129
Standardized Final Score (GRADE 8)	0.115	0.142	0.921	211

Notes: There are three separate panels in the table, each containing standardized scores on the pre-test and post-test. Standardization was performed within each panel-test grouping, excluding scores for students who missed the final exam. For example, pre-test scores for the quota treatment were standardized using the mean and standard deviation of pre-test scores for all 7^{th} and 8^{th} graders in the quota treatment who took both the pre-test and post-test.

labor-supply decisions in response to AA. Our experiment involves relatively low value prizes and a relatively low-stakes exam; yet we still find sizeable and significant effects of AA on student motivation and test performance.

From a policy perspective, these findings are important, as they indicate how AA not only promotes more racial diversity on college campuses, but at the same time it may also narrow achievement gaps between Whites/Asians and Blacks/Hispanics by

motivating higher levels of pre-college human capital investment on the part of under-represented minority students.

References

Peter Arcidiacono. Affirmative Action in Higher Education: How do Admission and Financial Aid Rules Affect Future Earnings? *Econometrica*, 73(5):1477–1524, 2005.

Eric Bettinger. Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores. *Review of Economics and Statistics*, 94:686–698, 2012.

Aaron Bodoh-Creed and Brent R. Hickman. Using Auction Theory to Study Human Capital Investment in Assortative Matching Markets: a Look at Affirmative Action in College Admissions. *Typescript, University of Chicago Department of Economics*, 2014.

William G. Bowen and Derek Bok. *The Shape of the River: Long-Term Consequences of Considering Race in College and University Admissions*. Princeton, NJ: Princeton University Press, 1998.

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

David L. Chambers, Timothy T. Clydesdale, William C. Kidder, and Richard O. Lempert. The Real Impact of Eliminating Affirmative Action in American Law Schools: An Empirical Critique of Richard Sander's Study. *Stanford Law Review*, 57:1855–1897, 2005.

Christopher Cotton, Frank McIntyre, and Joseph Price. Gender differences in repeated competition: Evidence from school math contests. *Journal of Economic Behavior and Organization*, 86:52–66, 2013.

Bruno Ferman and Juliano Assuncao. Does Affirmative Action Enhance or Undercut Investment Incentives? Evidence from Quotas in Brazilian Public Universities. *Typescript, Massachusetts Institute of Technology Department of Economics*, 2011.

Roland Fryer. Financial Incentives and Student Achievement: Evidence From Randomized Trials. *Quarterly Journal of Economics*, 126:1755–1798, 2011.

Brent R. Hickman. Human Capital Investment and Affirmative Action: A Structural Policy Analysis of US College Admissions. *Typescript, University of Chicago Department of Economics*, 2013.

Jessica S. Howell. Assessing the Impact of Eliminating Affirmative Action in Higher Education. *Journal of Labor Economics*, 28(1):113–66, 2010.

Michael Kremer, Edward Miguel, and Rebecca Thornton. Incentives to Learn. *Review of Economics and Statistics*, 91:437–456, 2009.

Edwin Leuven, Hessel Oosterbeek, and Bas van der Klaauw. The effect of financial rewards on students' achievement: Evidence from a randomized experiment. *Journal of the European Economic Association*, 8:1243–1265, 2010.

Mark C. Long. College Quality and Early Adult Outcomes. *Economics of Education Review*, 27:588–602, 2008.

Linda Datcher Loury and David Garman. Selectivity and Earnings. *Journal of Labor Economics*, 13(2):289–308, 1995.

Jesse Rothstein and Albert H. Yoon. Affirmative Action in Law School Admissions: What do Racial Preferences Do? *University of Chicago Law Review*, 75(2):649–714, 2008.

Richard H. Sander. A Systemic Analysis of Affirmative Action in American Law Schools. *Stanford Law Review*, 57:367–483, 2004.

Andrew Schotter and Keith Weigelt. Asymmetric Tournaments, Equal Opportunity Laws and Affirmative Action: Some Experimental Results. *Quarterly Journal of Economics*, 107(2):511–539, 1992.

APPENDIX

6.1. **Time Truncation Rule.** Time on our website was measured at the page level for each attempt of a quiz by each student. Pages contain blocks of either 3, 4, or 5 questions, so we divided each block-level time observation by the number of questions in order to get a measure of time spent per question. One difficulty arose in that there were a small number of clear instances where students left the website in the middle of a quiz for several hours or more. For example, the largest recorded time spent on a single question was 2,801 minutes, or roughly 47 hours. In order to correct this problem, a small number of implausibly large time observations needed to be corrected. After selecting a truncation point on the time-per-question domain, we replaced each observation above that point with the student-specific censored mean of time per question. For example, suppose that Tommy attempted 11 questions with observed times of 5 minutes for the first five, 15 minutes for the next five, and 300 minutes for the last, and suppose that the truncation point were 30 minutes per question. Then the eleventh observation of 300 minutes is replaced by Tommy's idiosyncratic censored mean of 10 minutes.

In order to select an appropriate truncation point we looked for occurrences of "holes" in the support of the distribution of times per question, or in other words, points at which a full support condition fails. We began with a natural assumption on the student type distribution that there are no interval subsets of the support interior where the type density assigns zero mass to the entire interval. If this condition holds, then since time spent on a question is a continuous choice related to one's type, that distribution should also have full support too. That is, unless some observations reflect time elapsed outside of learning activity, say due to work stoppages in the middle of a quiz. Thus, a straightforward way to search for spurious time observations is to sort the data and look for points at which a kernel smoothed density estimate (KDE) equals zero for some

interval of positive length. This idea gives rise to the following data-driven algorithm for selecting a truncation point:

(1) Sort all time observations from least to greatest, so that the j^{th} and $(j+1)^{\text{st}}$ observations are ordered by $t_j < t_{j+1}$ for all j.

(2) Using the sample $\{t_j\}_{j=1}^{J}$, compute an appropriately chosen bandwidth h_1 for a KDE based on a kernel function with support on $[-1,1]$.[29] Then find the smallest $j_1^* < J$ such that $t_{j_1^*+1} - t_{j_1^*} > 2h_1$. If no such j_1^* exists, then stop; no truncation is needed.

(3) Define initial truncation point $\tau_1 \equiv t_{j_1^*} + h_1$, and compute bandwidth h_2 for the KDE based on the censored sample $\{t_j\}_{j=1}^{j_1^*}$.

(4) In each subsequent iteration $k = 2, 3, \ldots$, if there exists j_k^* defined by

$$j_k^* \equiv \min\{j : t_{j+1} - t_j > 2h_k; j < j_{k-1}^*\},$$

then update the truncation point by $\tau_k \equiv t_{j_k^*} + h_k$, and re-compute bandwidth h_{k+1} for the KDE based on the censored sample $\{t_j\}_{j=1}^{j_k^*}$.

(5) Stop once k is found such that j_k^* does not exist (meaning that for the censored sample $\{t_j\}_{j=1}^{j_{k-1}^*}$ a KDE is strictly positive everywhere).

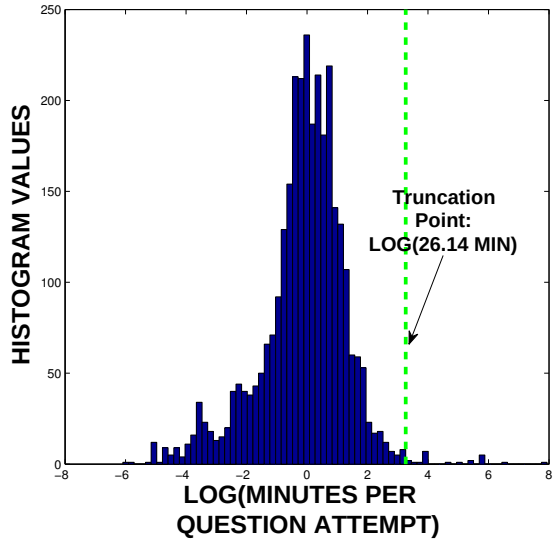
We chose a KDE based on the Epanechnikov kernel, which is known to be marginally more efficient than other kernel functions. This choice, in combination with Silverman's automatic bandwidth selection rule, implies a bandwidth formula of $h_1 = 2.345 S_1 J^{-1/5}$ in the first iteration, and $h_k = 2.345 S_k (j_{k-1}^*)^{-1/5}$ in the k^{th} iteration ($k \geq 2$), where S_k is the sample standard deviation within the k^{th} iteration. Notice that the algorithm does not actually require computation of a KDE at each iteration, only a bandwidth.

Executing this process on our data leads to a final truncation point of $\tau_2 = 27.81$ minutes per question (the 99.35^{th} percentile of the un-censored sample), after 2 iterations. Figure 7 displays a histogram of time spent per question, including observations above and below the truncation point. Time units are depicted in logs rather than levels for ease of visualization since the largest and smallest observations differ by several orders of magnitude.

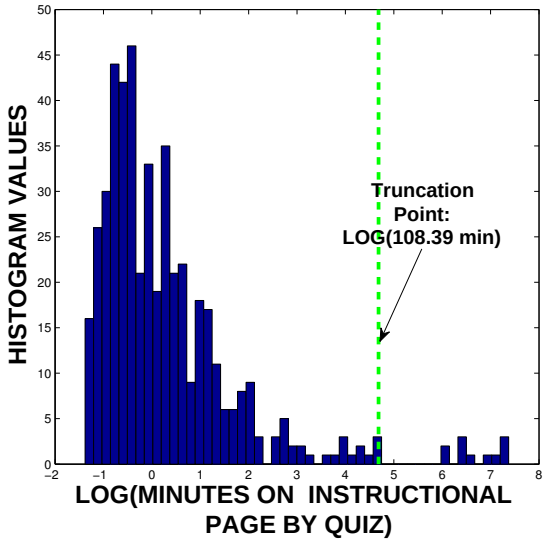
Table 6 displays the mean and variance of time spent per question attempt, using the censored sample of non-truncated times. Some subjects appeared more challenging in terms of the time students took to solve problems. The most difficult subject was

[29] Actually, the only crucial condition here is that the kernel function have bounded support. For example, in this context a Gaussian kernel would not do, as it places positive mass on the entire real line for any dataset. This would be equivalent to assuming full support *ex ante*.

FIGURE 7. TIME TRUNCATION RULE



(A) This panel displays a histogram of observed time spent on each question. Each datum in the histogram is a student-question-attempt observation.



(B) This panel displays a histogram of time per instructional page view. Each datum in the histogram is a student-quiz-attempt observation.

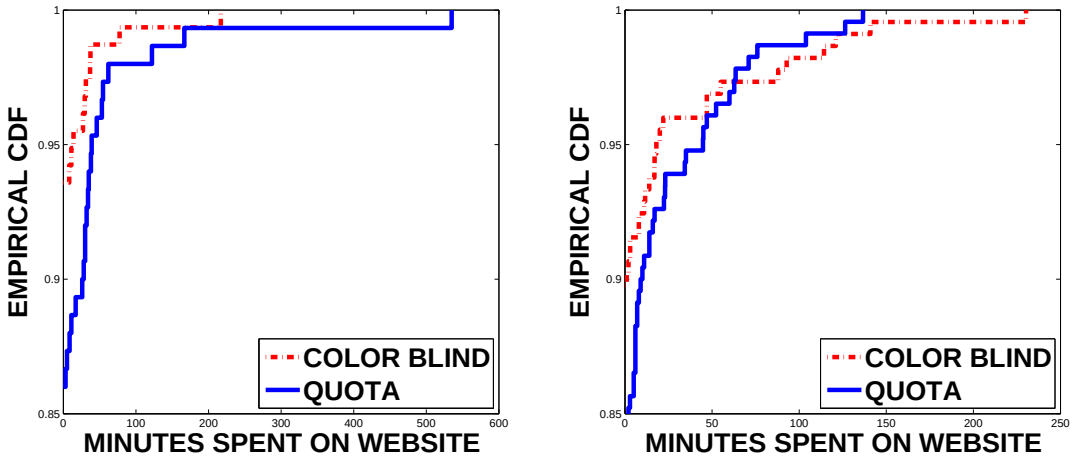
TABLE 6. Time Per Question by Subject

Subject	**Censored Mean Minutes Per Question**	**Std. Dev. (minutes)**	Mean + 2×Std. Dev.
Algebra	2.422	2.563	7.548
Arithmetic	1.398	1.238	3.874
Combinatorics	2.839	3.532	9.903
Geometry	2.183	2.577	7.337
Logic	1.807	1.742	5.291
Probability	1.996	1.137	4.27

combinatorics, with a mean time per question attempt of 2.839 minutes and a standard deviation of 3.532. The least difficult subject appeared to be arithmetic, with mean and standard deviation of 1.398 and 1.238, respectively.

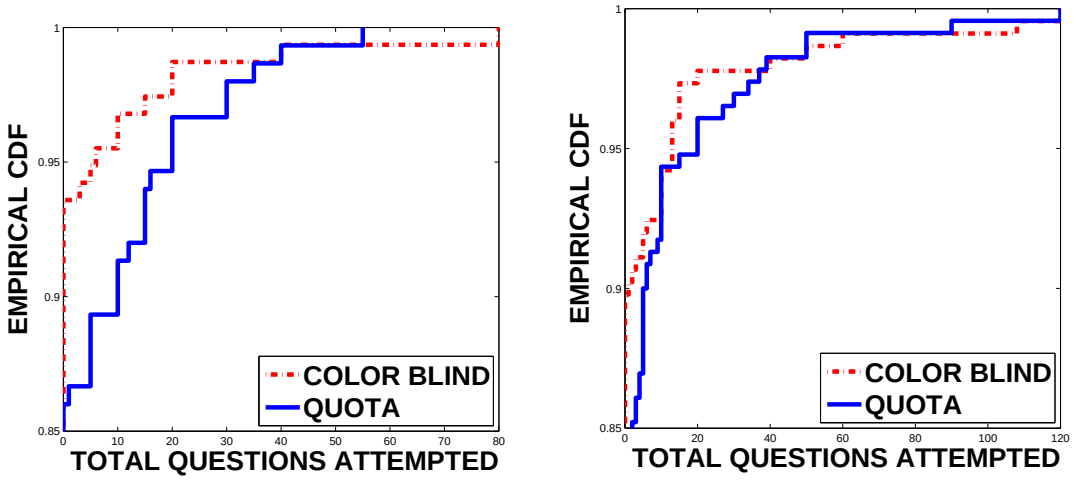
6.2. **Additional Figures.** Here we present some additional figures depicting the empirical distributions of investment activities by group and treatement status. In interpreting these figures, one caveat should be kept in mind. The theoretical predictions of Bodoh-Creed and Hickman [2014] only directly apply to the plots in Figures 5 – 6, since these

FIGURE 8. EMPIRICAL INVESTMENT DISTRIBUTIONS: COLOR-BLIND VS QUOTA



(A) Seventh Graders (B) Eighth Graders

FIGURE 9. EMPIRICAL INVESTMENT DISTRIBUTIONS: COLOR-BLIND VS QUOTA



(A) Seventh Graders (B) Eighth Graders

depict CDFs of exam scores, the variable being directly incentivized within the experimental study. Thus, theory predicts that those plots should qualitatively resemble the patterns in Figure 3. It has nothing directly to say about other intermediate variables such as time spent on the website, or number of questions attempted, as these may combine in different ways for different agents to produce exam scores. However, for illustrative purposes, we present additional CDF plots here.

FIGURE 10. SEVENTH GRADE TREATMENT GROUPS:
PRE-TEST VS FINAL EXAM

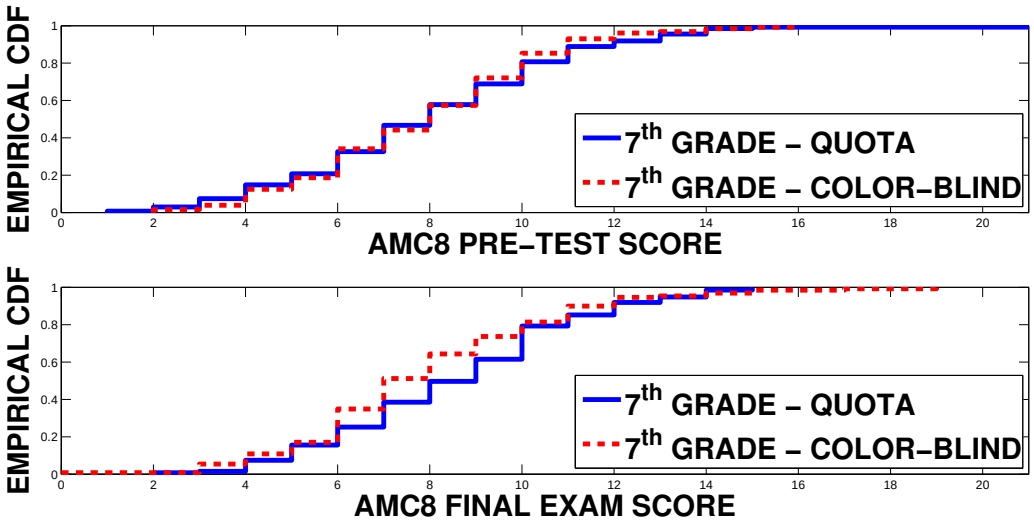


FIGURE 11. EIGHTH GRADE TREATMENT GROUPS:
PRE-TEST VS FINAL EXAM