Affirmative Action and Human Capital Investment: Evidence from a Randomized Field Experiment

Christopher S. Cotton, Brent R. Hickman, and Joseph P. Price

Abstract. We conduct a field experiment paying students based on relative performance on a mathematics exam and tracking study efforts on a mathematics website to test the incentive effects of Affirmative Action (AA) policies on study effort and math proficiency. AA increases study effort and exam performance for the majority of disadvantaged students targeted by the policy. While the performance of the highest-ability students targeted by the AA policy declines, on average study activity and exam performance rise under AA. Overall, the experimental evidence suggests that AA can promote greater equality of market outcomes while narrowing achievement gaps.

Cotton: Department of Economics, Queen's University, Dunning Hall 230, Kingston, Ontario K7L 3N6; cotton@econ.queensu.ca; phone: (613) 533-2251.

Hickman (corresponding author): Olin Business School, Washington University in St. Louis; One Brookings Drive Campus Box #1133, St. Louis, MO 63130; hickmanbr@gmail.com .

Price: Department of Economics, Brigham Young University, 162 FOB, Provo, UT 84602; joe_price@byu.edu; phone: (801) 422-5296.

Original version: October 2013; This version: August 2020. We greatly appreciate comments and helpful input from Andrew Sweeting, Aaron Bodoh-Creed, John-Eric Humphreys, John List, and participants of seminars and conference at the University of Chicago, the American Economic Association meetings, the North American summer meetings of the Econometric Society, the Canadian Public Economic Group conference, the Econometric Society World Congress, the European Meetings of the Economic Science Association, Brigham Young University, University of Missouri-Columbia, Queen's University, Arizona State Univeristy, Ryerson University, SUNY Binghampton, UBC-Okanagan, University of Guelph, Simon Fraiser University, and London School of Economics. We also wish to acknowledge the outstanding research assistance of Joe Patten, who played a crucial role in executing this project. Cotton is grateful for the financial support provided by his position as the Jarislowsky-Deutsch Chair in Economic and Financial Policy at Queen's University. Cotton and Price gratefully acknowledge partial funding from the Spencer Foundation which supported an initial pilot study that led to this research. An initial draft of this paper appeared as NBER working paper 20397 in August 2014.

NOTE: previous versions of this manuscript have circulated under the title, "Incentive Provision in Investment Contests: Theory and Evidence."

1. Introduction

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

In the context of American university admissions, the rationale for AA does not stem from concerns over proactive discrimination; rather, it is that the university market is a rank-order academic competition where some demographic groups are at a fundamental disadvantage to other groups due to residual effects of past institutionalized discrimination. That is, African-American and Hispanic children on average grow up in homes with less wealth and income, attend lower-performing K-12 schools, have less-educated parents and extended-families, and have less access to other developmental inputs such as healthcare and tutoring.[2] In turn, AA attempts to compensate for the competitive disadvantage that may be caused by historic and ongoing inequality by giving special consideration for race when judging college applications. There is a substantial empirical literature studying the effects of AA, focusing mainly on its direct impact at the point when university admissions outcomes are determined.[3] However, this literature has largely treated pre-college academic achievement as exogenous and mechanically determined by a student's innate characteristics and environment. It has thus ignored the implications of AA for student behavior *prior* to the admissions process when they choose their effort level for human capital development, which in turn determines the grades and exam scores that will eventually go on their college applications.

AA policies change the set of admissions offers that a student may perceive as attainable at the beginning of the competition—e.g., before/during high school, when goals are set and effort choices made—and which other students he/she must beat out in competing for these offers. These changes in the competition may alter students' incentives to prepare for college as they

[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).

[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'... Thus it is not enough just to open the gates of opportunity. All our citizens must have the ability to walk through those gates... 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 et al. (2005), have estimated the impact of AA on graduation rates among minority groups.

study to improve grades and test scores or work to otherwise increase their human capital. In theoretical work complementing this paper, Bodoh-Creed and Hickman (2018) and Cotton, Hickman and Price (2020) investigate these incentives, showing that the impact of AA on effort and achievement depends on one's *relative* academic ability and demographic group. Cotton, Hickman and Price (2020) explore comparative static predictions that produce three insights relevant to inequality and competitive college admissions. First, holding fixed one's own ability and the intrinsic costs and benefits of real skill formation, a student's choice of how much productive human capital to develop changes when the distribution of rivals' abilities changes, leading to shifts in the intensity of competition for college seats. This first result refutes traditional wisdom that high-school grades and exam scores are exogenous or mechanically determined. Second, they show that the sign and magnitude of one's response to AA depend on one's position in the ability spectrum. A student who benefits from an AA policy may increase investment if they now have a shot at winning admission to a prestigious school that was otherwise out of reach. Some beneficiaries at the top of the ability distribution may decrease investment if AA does not meaningfully improve their likely college placement but makes it easier to attain. AA leads to (generally weaker) responses among non-beneficiaries may increase or decrease investment depending on where they rank among competitors. Third, Cotton, Hickman and Price (2020) show that some forms of AA may actually increase the aggregate stock of human capital attainment while reducing the average achievement gap between disadvantaged demographics and their competitors.

If the sophisticated strategic behavior in their model is a meaningful reflection of real-world decision-making among future college applicants, then it has important implications for ongoing debates among both academics and policy-makers. The current paper explores the empirical validity of these predictions using field experimental methods to represent salient features of competitive human capital investment running up to the college admissions market. We execute a large, school-based math competition with an intermediate human capital investment period and many heterogeneous prizes to study how an AA policy benefiting a disadvantaged group changes students' study effort and test performance. On a general level, our experiment with hundreds of heterogeneous students competing for heterogeneous prizes is the first controlled experiment of a large-scale contest environment. On a more practical level, it provides experimental evidence that students respond to AA incentives in a manner that is highly consistent with predictions of recent economic theory.[4]

In our experimental design, cash prizes stand in for university admission outcomes to motivate test subjects to spend leisure time learning math.[5] Our experiment involves paying middle-school-aged students based on their relative performance on a mathematics exam, in a similar

[4]Earlier applied theory on AA incentives includes Coate and Loury (1993) and Moro and Norman (2003). Our experiment with many-to-many matching, demographic cost asymmetry, and scarcity of high-value prizes, is a better representation of the more-recent work by Bodoh-Creed and Hickman (2018) and Cotton, Hickman and Price (2020).

[5]Our experiment covers short-run incentives on the scale of weeks. Real-world pre-college investment spans years, but it is not feasible for the researcher to experimentally create exogenous identifying variation on that scale. This paper uses feasible variation in short-run incentives to focus on how AA shapes competitive behavior driving individual labor-leisure decisions and human capital accumulation. See Section 5.1.2 for further discussion.

fashion as outcomes are determined on the university matching market using college entrance exams and high school grades. In order to create a clean test of theory on investment and discouragement effects, we use grade cohort as our demographic delimiter. This distinction mirrors some important aspects of racial disparities—our *disadvantaged group* (*e.g.*, 7^{th} graders) have received, on average, fewer inputs into their math education relative to their *advantaged group* counterparts (*e.g.*, 8^{th} graders), but with substantial overlap in the ability distribution notwithstanding—while filtering out other factors such as cultural differences or stereotype threat that could confound the effects we seek to investigate.

Students are individually randomized into one of two treatments: a pure-rank-order treatment in which older and younger students compete for the same set of prizes, and an AA treatment (i.e., a representative quota) in which a representative set of prizes is reserved for the younger students before the competition begins.[6] In other words, AA reserves a more-favorable set of prizes for students in the lower grade, relative to what they receive (in equilibrium) when competing directly against higher-grade students. The representative quota has the opposite impact for higher-grade students, who now receive a less favorable distribution of prizes. We find heterogeneous treatment effects that conform well to theory: some individuals perform better and others worse under the AA policy, while the average performance of disadvantaged students increases and achievement gaps between groups decline.

Because we are interested in how AA changes incentives to invest real effort into developing human capital, a typical classroom experiment in which students are assigned to a treatment immediately before completing some task is not ideal. Instead, we worked with teachers to incorporate the experiment into their students' schooling over a course of two weeks. We assessed two in-class mathematics exams separated by 2 weeks—a pre-exam to measure baseline proficiency and a post-exam to allocate prizes and measure progress—and we provided the students with a 10-day interim period in which they voluntarily could study math problems ahead of the final exam. During this period we provided the students access to a website with practice materials and tracked their use of this site.

Some previous experimental studies have investigated the link between AA and effort using head-to-head competition (e.g., Calsamiglia, Franke and Rey-Biel (2013)). There are several important differences between our research design and previous work. First, our experiments involve many-to-many rank-order matching in order to capture various salient features of real-world examples like college admissions. The size of our contests allows us to estimate differing signs and magnitudes of quantile-specific behavioral shifts. Second, ours is also the first experimental paper to track interim competitive investment in human capital, rather than only same-day effort. By showing that students within the same grade cohort (but in different treatments) access the practice website at different rates, ours is the first experimental study to identify

[6]We focus on a representative quota form of AA for its analytic and experimental tractability. Alternative forms of AA involve members of $\mathcal{D}$ receiving an achievement bonus instead. Bodoh-Creed and Hickman (2018) define broad classes of bonus systems and quota systems (of which RQ is a special case) and prove the strategic equivalence between the two classes of mechanisms.

an effect of AA incentives on observable study effort. Third, we worked with students' regular teachers, using materials that were already part of the curriculum in their schools, and allowing for autonomous labor-leisure choices at home. Our goal was to create a natural learning environment, where the free-time/study-time trade-offs were familiar to test subjects, and thus to mimic scenarios they face when preparing for their future involvement in the actual university admissions process.

We find strong evidence that AA influences labor-leisure trade-offs and measured proficiency by our disadvantaged group. Although the highest-ability students from the disadvantaged group decrease performance on the exam, the majority of disadvantaged students increased study time and test performance under AA. On average, the AA policy increased test scores among the disadvantaged group by a fifth of a standard deviation over the course of the experiment. While some portion of this increase may be due to greater effort/focus on the day of the exam, we also show that disadvantaged students in the AA treatment are more than twice as likely to use the practice website to prepare for the exam, spend nearly triple the amount of total time, and practice questions from more math topics relative to their counterparts in the non-AA treatment. One might worry that these gains come at the cost of commensurate weakening in incentives for students who do not benefit from AA, but the data do not support this concern. We find only weak evidence of behavioral changes among the advantaged group, leaning toward a small (though insignificant) increase in average investment. A further analysis also suggests that AA can help narrow demographic achievement gaps by incentivizing increased study effort among disadvantaged students.

It is important to acknowledge that our experiment studies AA while abstracting from various contextual factors that make AA such a complex and challenging issue. Our experiment relies on a fairly homogeneous study population, defining one's disadvantaged status based on grade level, rather than on race, ethnicity, gender, income, or other factors that correlate with real-world inequality. On one hand, our sample population enables a clean experimental design for testing whether students plausibly respond to incentives in accordance with equilibrium predictions. On the other hand, since our study does not capture many salient socioeconomic and cultural features of real-world environments where AA is used, it may not capture the full impact of such policies on student behavior and performance gaps in practice. Section 5.1 below discusses real-world interpretations and limitations of our work in further detail.

Our paper forms part of a nascent literature on how AA in admissions affects academic performance ahead of the application process, including Bodoh-Creed and Hickman (2019) and Akhtari, Bau and Laliberté (2020) based on observational data. The former identifies structural counterfactuals and the latter tests for behavioral responses to changes in admissions criteria. Our field experimental study finds evidence consistent with these two studies and contributes to the empirical discussion in two ways. First, we directly investigate the validity of the behavioral assumptions on which these other papers base their causal estimates. Second, we investigate the link between changes in intermediate inputs (through study-time tracking on our website) and changes in measured academic proficiency.

This project is also related to work on performance pay in schooling. Recent studies including Burgess, Metcalfe and Sadoff (2016), Bettinger (2012), Fryer (2011), and Leuven, Oosterbeek and van der Klaauw (2010) found mixed results about whether paying students can improve their effort and outcomes. Our experimental incentives are different from these in that we paid participants based on *relative performance*, rather than offering a fixed-wage contract. We found that test subjects who engaged in home study in response to our incentives earned a relatively high expected wage of $10.73 per hour of study, which may help explain the strength of the behavioral response we found. Kremer, Miguel and Thornton (2009) also studied a program awarding merit scholarships based on relative performance. Our experimental design differs from these studies in that we were able to monitor outcomes *and* interim time inputs, and our focus is not on the effectiveness of pay-for-performance *per se*, but rather on differences between AA and pure-rank-order incentives.

The remainder of this paper is as follows. Section 2 provides an intuitive summary of the testable predictions from Cotton, Hickman and Price (2020) which underpin our experimental design. Section 3 outlines our field experiment while Section 4 presents empirical results. Section 5 discusses our findings and concludes. An online appendix contains additional tables, graphs, and details.

2. TESTABLE PREDICTIONS FROM ECONOMIC THEORY

Cotton, Hickman and Price (2020) provide a model of competition among students for post-graduation opportunities such as college admissions, scholarships, or jobs, which we refer to generally as *prizes*. Students choose time and effort to invest in building human capital (HC), but individuals differ in their achievement costs. For example, students vary by underlying ability, access to efficient study resources, or opportunity cost of time. Below we use the descriptors *high aptitude* and *low achievement cost* interchangeably, since costs of and aptitude for producing HC are inversely related. Prizes are allocated based on relative HC achievement among all competing students. The model is an all-pay contest with many heterogeneous students simultaneously competing for many heterogeneous prizes, and each student may consume only one college seat/job/etc.

The model is designed to assess how the distribution of endogenous HC investment depends on AA policy. There are two observable demographic groups, $\mathcal{A}$ and $\mathcal{D}$. Although individuals within each group vary by aptitude, members of group $\mathcal{A}$ find building HC less costly, on average, than members of group $\mathcal{D}$.[7] In the baseline analysis, referred to as the pure rank order (PRO) model, students who accumulate more HC get better post-graduation prizes, regardless of which group they are from. In the college context, this is often referred to as "color-blind" admissions. We also consider an AA policy in which a representative set of prizes is set aside for group $\mathcal{D}$ before competition begins; we refer to this as the representative quota (RQ) model.

[7]Specifically, the quantiles of the cost distribution in group $\mathcal{D}$ are assumed to be higher than the corresponding quantiles of $\mathcal{A}$ so that the same aptitude spectrum exists in both groups, but the relative frequencies of high and low costs are different. In real-world competitive HC investments, relative disadvantages arise from asymmetric resources such as differing wealth/income, K-12 school quality, or access to fundamental inputs such as childhood healthcare. In many societies, these disadvantages are systemically tied to race, caste or ethnicity for a variety of historical reasons.

Since the RQ regime reserves a proportional set of prizes for each group, from students' perspectives the distribution of prize quality is the same as under PRO competition, and the meaningful difference between the two involves a shift in the distribution of competitors they each face. For any student, the AA policy may make the competition more or less intense, by pitting her against other students with higher or lower aptitudes, and therefore by changing the set of post-graduation outcomes that are feasibly within reach. Under an increase in the intensity of competition, students adjust their optimal labor-leisure calculation and develop more or less HC in response. The model implies three testable predictions relating to HC investment with or without AA. All three involve a comparative static scenario where the overall distribution of prizes and their intrinsic value are held fixed, but the intensity of competition changes.

Prediction $(\mathbf{I})$: Suppose that competition becomes more fierce in the sense that a given person with fixed achievement costs faces a distribution of competitor costs where the quantiles are all lower than before. Then if she is among the highest aptitude students she will react by increasing HC investment, while middle- and low-aptitude students reduce investment due to a discouragement effect as better prizes are now perceived as out of reach.

Prediction $(\mathbf{II})$: Replacing a PRO rule with a RQ rule brings more valuable prizes within reach of low- and middle-aptitude members of $\mathcal{D}$, thus mitigating discouragement effects and increasing their HC investment choice. The best and brightest members of $\mathcal{D}$ invest less in HC under RQ because they now face less intense competition for the top prizes. The effects on members of group $\mathcal{A}$ from each ability level tend in opposite directions by similar logic.

Prediction $(\mathbf{III})$: In the aggregate, a RQ rule will lead to gains in achievement for most members of group $\mathcal{D}$, relative to a PRO rule. Among group $\mathcal{A}$ behavioral responses are generally weaker in magnitude and the theory is ambiguous about the sign of the majority response.

Together, these three predictions have important implications for the impact of AA on inequality and aggregate investment in productive HC. A RQ policy will decrease the average achievement gap between groups $\mathcal{D}$ and $\mathcal{A}$ as most students in the former group increase effort, and it may even raise the overall stock of HC. However, in order to argue that these implications are empirically relevant one must have confidence that real-world students preparing for college are capable of behaving in a way consistent with the sophisticated strategic behavior depicted within the model. Therefore, a key aspect of our field experiment is to engineer the counterfactual scenarios described above, where same-ability students face differing levels of competition for the same set of outcomes in a real-effort learning exercise. We can then test whether the RQ policy induces the predicted changes in final achievement, and we can also probe for changes in the intermediate inputs behind investment costs: time and effort.

3. EXPERIMENTAL DESIGN

3.1. **SAMPLE POPULATION.** We built our incentives and learning exercise around the American Mathematics Competition 8 (AMC8) exam, sponsored by the Mathematical Association of America, for students in 8^{th} grade and below. It consists of 25 multiple choice questions (5 choices each) in 40 minutes, and the questions become progressively more difficult from start to finish.

The MAA explains that it "provides an opportunity to apply concepts [to] high-level problems which... challenge and offer problem-solving experiences beyond those provided in... school."

Our sample includes 992 suburban middle-school and junior-high students from 10 schools in Utah County, Utah, including charter schools and regular public schools. School administrators partnered with us for the study, so participation was at the classroom level on an opt-out basis. Most of our partner schools had previously participated in the AMC8 before partnering with us for this study. Academically and socioeconomically, our sample population was fairly average within the US.[8] One difference between our sample population and the national population was a greater degree of homogeneity: roughly 9% of subjects were racial minorities (black or Hispanic), compared to a nationwide average of 37%.

3.2. **TREATMENT GROUPS AND INCENTIVES.** Participants in our study first took a practice AMC8 test from a previous year. We used this as a baseline measure of each subject's math proficiency. Individuals were randomized into either a control group—with a pure rank order (*PRO*) competition—or a treatment group—with a representative quota (*RQ*) competition. For the *PRO* group we ran competitions involving subjects in two adjacent grades; that is, 7^{th} and 8^{th} graders competing together, and 5^{th} and 6^{th} graders competing together. For each of the two age cohort pairings, subjects in the lower grades (5^{th} or 7^{th}) are the "disadvantaged" group $\mathcal{D}$, and subjects in the higher grades (6^{th} or 8^{th}) are referred to as the "advantaged" group $\mathcal{A}$, since the latter are one year older and have received one more year of mathematics education on average. For treatment *RQ*, subjects competed only within their own grade level, but for a proportionally equivalent set of prizes (relative to *PRO*), as described below. We ran separate RQ competitions for 5^{th}, 6^{th}, 7^{th}, and 8^{th} graders.

The top 30% of subjects within each competition group received cash prizes, which were uniformly distributed between \$4 and \$34 in \$2 increments. Prizes were awarded within competition groups, according to final exam scores. For example, 7^{th} grade subjects in the *PRO* competition needed to score within the top 30 percent of all 7^{th} and 8^{th} grade subjects in order to receive a prize. In treatment *RQ*, subjects competed against others in their own grade only, but for a representative set of prizes. More specifically, we began with the same aggregate prize distribution as for treatment *PRO*, and then earmarked prizes at each different level in proportion to the mass of lower-grade subjects in each age-cohort pairing (the masses of groups $\mathcal{A}$ and $\mathcal{D}$ were also identical across treatments *PRO* and *RQ*). This ensured that the moments of the prize distribution (including the 70% mass of zeros) were the same across all competition groups, with each one vying for the same number and variety of prizes on a per capita basis. For example, 7^{th} grade subjects in treatment *RQ* only had to score within the top 30% of 7^{th} graders in their treatment

[8]Our test subjects were somewhat more affluent but academically comparable to the rest of the country. In 2012, 33% were eligible for free or reduced-price lunch, compared to a national average of 48%. Our partner schools housing 5^{th} and 6^{th} graders (20% of our sample) performed significantly better than other Utah schools at meeting state math standards (approximately 91% vs 76%), while our schools housing 7^{th} and 8^{th} graders (80% of our sample) performed slightly worse (81% vs 83% meeting state standards). Utah ranked at or near the median for nationally measured academic outcomes such as NAEP scores and enrollment rates in Advanced Placement programs.

to receive a prize. Moreover, their distribution of prizes, conditional on winning something, was the same as for 8^{th} graders in treatment *RQ*, and also the same as for all $7^{th}/8^{th}$ grade subjects in treatment *PRO*. Thus, for a subject of a given ability level in either group, the only difference across the two treatments is the distribution of one's competitors.

Each subject received an information sheet describing their assigned group, how many subjects from which grade(s) they would be competing against, and the score distribution within their group based on the practice test. Subjects received their own practice score back at the same time so they could see where they fit within their competition group. The sheet also contained a table describing the prize structure. We printed information relative to each competition group on a different color of paper so that subjects could visually see in their classroom that roughly half of the subjects 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 $5^{th}/6^{th}$ grade and one for $7^{th}/8^{th}$ grade). In Online Appendix C we provide an example of the information sheet given to each group.

3.3. **MATH LEARNING WEBSITE.** The information sheet provided a URL to a website we set up with 125 practice problems (drawn from 5 past AMC8 exams) covering six different math topics: 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 problems 11-20, and a third covering problems 21-25. Test subjects were notified that each grouping of 3 same-year quizzes was ordered by their difficulty level. We also arranged this same battery of math problems into an additional set of 16 quizzes, each containing 5 topic-specific math problems. These topic quizzes were also ordered by their difficulty level.

Subjects could attempt each quiz multiple times if they wished. After completing each quiz, our software displayed an instructional page that reported to each subject her score, the correct answers for each problem, and step-by-step solutions published by the developers of the AMC8. In order to access the website that contained links for all practice quizzes, students had to input their name, grade, and school into the web form. This allowed us to track online activities for each quiz session, including which students visited the website, how many different math topics 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.

3.3.1. ***Time Measurement.*** 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. Time was measured at the page-view level, meaning that we got a time measure for blocks of either 3, 4, or 5 questions. In order to convert this time information into a per-question measure, we divided each block-level time observation by the number of questions within that block. One difficulty was instances where subjects 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 subject-specific censored mean. In selecting our truncation point we looked for occurrences of "holes" in the support of the distribution of times per question.[9,10] For our time per question data, this leads to a truncation point of 26.14 min/question (the 99.35th percentile), and for instructional page views, 108.39 min/page view (*i.e.*, 21.68 min/solution, or the 98th percentile). In Online Appendix B.3 we display a histogram of (uncensored) time per question and instructional page view times.

At the end of the day, it is impossible to directly observe work stoppages in the middle of a quiz question, and we are effectively interpreting work stoppages of less than 26 minutes as time which comes at a positive cost to the child. We argue that 26 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 subjects. By this metric, the most difficult topic was 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 98th 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 naturally exist on the interval between 10 minutes and 26 minutes.

Before moving on, a final concern about time use monitoring is worth discussing. Given our experimental setup which monitors time use through our AMC8 practice website, it is impossible for us to guarantee that test subjects did not spend time off of our website studying math in preparation for the incentivized exam on their own. We believe this is not a major concern for our empirical results for several reasons. First, the typical AMC8 problem is perceptively different from the typical math problems contained in a child's standard textbook that he/she would have as an outside resource. AMC8 questions focus more on creative problem solving using age-appropriate math tools, while textbook problems focus on repetitive learning by doing. Since our students took an AMC8 pre-test before receiving access to the practice website, they would likely have noticed this difference. Second, although other online resources were available at the time (e.g., Khan Academy), our information sheets emphasized that the most relevant study materials could be found on our website, which would be stocked with "practice problems that we gathered from past years' exams of the AMC8." Third, because of randomization, the same set of outside study options would have been available to all students in all groups and treatments, on average. To the extent that students studied for the exam outside of our website

[9]A hole was defined as the minimum interior point at which a full-support condition fails (i.e., a point below the sample maximum where a kernel-smoothed density estimate hits zero). See Online Appendix B.3 for further details.

[10]To illustrate, suppose Tommy attempted 15 questions, with an average of 10 minutes on the first 14, but 2,000 minutes recorded for the 15th because he left the site before finishing. Then the last observation is replaced by Tommy's censored mean of 10 minutes. We interpolate the datum in this situation (rather than dropping it) because from the website log we know Tommy spent time on the 15th problem; we aren't sure how long, but on average the missing time observation should be the same as his censored mean. Dropping these observations instead produces very similar results.

then, our experimental data would understate the differences between our control and treatment groups. Finally, this concern applies only to our measurements of inputs (i.e., time and study activities), but does not apply to our measurements of final outputs (i.e., exam scores). Given that the measured changes in the latter seem to agree with measured changes in the former—that is, students whose study intensity on our website increased by most had AMC8 scores that increased most as well—we conclude that outside, non-monitored study activities are not a significant threat to the identification of treatment effects.

3.4. **TESTING.** Subjects 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 subjects in our study attended schools where participation in the AMC8 was already being offered to students by their teachers. Schools that cooperated in this study administered the test to all students in each participating classroom on an opt-out basis, so that all subjects participated in the study, except a small number whose parents returned an opt-out form. The practice pre-test was the AMC8 exam from the previous year, and the final exam was the AMC8 for the current year. The cash prizes were delivered to each school shortly after the final exam and handed out to each subject 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 using subjects' scores on the AMC8.

4. EXPERIMENTAL RESULTS

The theoretical model predicts that some subjects in each demographic group will increase their HC investment activities under a representative quota AA rule, while others in the same group will decrease their efforts. Our empirical investigation takes the next step by estimating magnitudes and testing for the signs of average effects of AA on effort and performance by demographic group. We also execute a smoothed, nonparametric CDF estimator to investigate heterogeneous treatment effects by quantiles and examine impacts on demographic achievement gaps.

4.1. **DESCRIPTIVE STATISTICS.** Table 1 contains descriptive statistics. Roughly one-quarter of our sample was $5^{th}/6^{th}$ graders that all came from accelerated classes, while the other three quarters were $7^{th}/8^{th}$ graders who were representative of the overall student body within their schools. This difference is born out in the data: while 8^{th} grade subjects did best on the pre-test with an average score of 9.04 (out of 25 possible), 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.[11]

We have also broken down test scores by two groups we refer to as *investors*—subjects who logged on to our website at least once—and *non-investors*—those who did not. Subjects who did better on the pre-test were more likely to be investors, although some who struggled on the pre-test were investors, and many who did quite well on the pre-test were non-investors. For the

[11]The national AMC8 population in 2013 (see https://amc-reg.maa.org/reports/generalreports.aspx) had a mean of 10.69 with standard deviation 4.44. The AMC8 is predominantly administered through opt-in participation, whereas our experiment was on an opt-out basis. This accounts for the lower mean among our sample pool.

TABLE 1. SUBJECT 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
Topics	1.94	1.00	1.43	118

Notes: All time figures are quoted in minute units. *Investors* are defined as subjects who logged on to the math learning website at least once during the investment period. *Non-Investors* are those who did not.

TABLE 2. EFFORT AND PERFORMANCE BY TREATMENT

	Investment				**Performance**
	Used Website	# Topics Attempted	Total Time	# Questions Attempted	Final Exam Score
Representative Quota	0.154	0.284	6.634	2.729	8.680
Std. Err.	(0.015)	(0.037)	(1.216)	(0.456)	(0.139)
Pure Rank Order	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.

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. If we divide total cash payments by total hours worked by all investors, we get an expected hourly wage of $10.73. The number of questions attempted ranged between 1 and 120, with mean and standard deviation of roughly 19 and 23, respectively. *Topics* represents the number of different topic categories a subject attempted, using the topic-specific quizzes, being about two on average.

4.2. **EMPIRICAL ANALYSIS.**

TABLE 3. TESTING DIFFERENCES BY TREATMENT

	Investment				Performance
	Used Website	# Topics 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	0.060***	0.081	2.495	0.793	0.154
P-Value:	*[0.003]*	*[0.112]*	*[0.143]*	*[0.212]*	*[0.375]*
(*Controls: pre-test scores, school FEs*)					
N	992	992	992	992	895

Notes: Each cell contains the coefficient for the quota treatment from a separate regression. Row 1 includes no controls and provides a test of the differences in Table 1. Row 2 includes a control for pre-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 HC investment for the total study population, and are therefore averaged over both investors and non-investors.

4.2.1. ***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 fraction of students from each treatment group who logged on to our website at least once to practice math. As for the other investment variables, the reader should keep in mind that Tables 2 - 4 aim to measure a treatment effect of a policy on an entire group, including 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. The results indicate that subjects in the quota treatment, including subjects from all age groups, were 75% more likely to have visited the website than subjects in the baseline *PRO* treatment. They also tried out more topics, spent more time on the website, and answered more questions. Table 2 also indicates that subjects in both treatments scored roughly the same on the final exam, with the point estimate being slightly higher for the quota treatment but not significantly so in a statistical sense. Later on we will see a different story when we separate these measures by 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 an *RQ* rule and *PRO* allocations at the population level (*i.e.*, including both groups $\mathcal{A}$ and $\mathcal{D}$). 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 there is no difference by treatment group. From the table we see strong evidence that the representative quota increases the fraction of all subjects willing to invest at least some time. We also see suggestive evidence that it induces them to experiment with more topics, as well as increase the total time invested and number of questions attempted. Although these last two differences are noisier (with p-vlaues of 0.107 and 0.146, respectively),

TABLE 4. TESTING DIFFERENCES BY DEMOGRAPHICS AND TREATMENT

	Investment				Performance	
	Used Website	# Topics Attempted	Total Time	# Questions Attempted	Final Exam Score	Exam Score Change
Constant ($\widehat{\beta}_0$)	0.079***	0.153***	3.312*	1.342*	8.087***	0.009
Std. Err.	(0.024)	(0.059)	(1.968)	(0.736)	(0.208)	(0.238)
Quota ($\widehat{\beta}_1$)	0.089***	0.151*	5.631**	1.333	0.614**	0.577*
Std. Err.	(0.033)	(0.083)	(2.754)	(1.030)	(0.287)	(0.332)
P-Value:	*[0.007]*	*[0.069]*	*[0.041]*	*[0.196]*	*[0.033]*	*[0.082]*
Advantaged $*$ *Quota* ($\widehat{\beta}_2$)	-0.047	-0.111	-5.082	-0.858	-0.701*	-0.547
Std. Err.	(0.042)	(0.105)	(3.504)	(1.310)	(0.360)	(0.416)
Advantaged ($\widehat{\beta}_3$)	0.017	0.065	1.109	0.814	0.580**	-0.179
Std. Err.	(0.030)	(0.076)	(2.513)	(0.940)	(0.261)	(0.297)
Pre-Test (*standardized*) ($\widehat{\beta}_4$)	0.032***	0.050*	1.016	0.624*	1.301***	*N/A*
Std. Err.	(0.011)	(0.027)	(0.886)	(0.331)	(0.092)	*N/A*
School Fixed Effects	yes	yes	yes	yes	yes	yes
N	992	992	992	992	895	895
Additional Test: Effect of Quota on Advantaged Group						
$\widehat{\beta}_1 + \widehat{\beta}_2$	0.042	0.040	0.549	0.475	-0.087	0.030**
P-Value:	*[0.103]*	*[0.545]*	*[0.800]*	*[0.557]*	*[0.691]*	*[0.905]*

Notes: Each column is a separate regression. *Advantaged* is an indicator variable for whether the subject is a 6^{th} or 8^{th} grader (the older group in each school type). We also include each subject's standardized pre-test score, where standardization is based on the mean and variance within each school type (*i.e.*, $5^{th}/6^{th}$ or $7^{th}/8^{th}$), in order to control for differences in starting HC. 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.

the estimated magnitudes are large, with RQ subjects logging an estimated 57% and 70% more inputs of time and question attempts, respectively. Additional controls (pre-test score and/or school fixed effects) are added in the bottom two rows as a check on the effectiveness of our randomization. Adding these variables caused no statistically detectable shifts in point estimates.

4.2.2. ***Testing Differences by Treatment within Demographic Groups***. Recall that the theory allows for AA to have differential effects by idiosyncratic ability (*i.e.*, achievement cost) and demographic group. In Table 4, we add a demographic dummy to investigate this claim. The first five columns each present estimates for a regression equation of the form

$$Outcome = \beta_0 + \beta_1 Quota + \beta_2 Advantaged * Quota + \beta_3 Advantaged + \beta_4 Pre\text{-}Test + \varepsilon,$$

where *Quota* is a dummy for treatment status, *Advantaged* is a demographic dummy, *Pre-Test* is a subject's pre-test score in standard deviation units, and the specific *Outcome* variable is

labeled in the column header.[12] With the inclusion of the interaction term $Advantaged * Quota$, the coefficient β_1 represents the average effect of the representative quota specifically on the disadvantaged group. The effect of the policy on the advantaged group is represented by the sum $\beta_1 + \beta_2$. The last column in the table moves pre-test score to the left-hand side to estimate the treatment effect on math improvement:

$$FinalExamScore - Pre\text{-}TestScore = \beta_0 + \beta_1 Quota + \beta_2 Advantaged * Quota + \beta_3 Advantaged + \varepsilon.$$

For completeness, all regressions include controls for school-level fixed effects.[13]

For members of the disadvantaged group, we find evidence of large, positive effects across all four investment measures. First, we see a highly significant 8.9 percentage point increase in disadvantaged subjects' willingness to spend at least some time on the website under AA. To put this in perspective, we can compute a within-demographic percent change for the disadvantaged group by $100 \times (\beta_1/\beta_0)\%$, which amounts to an increase of 113% on the extensive margin, relative to their disadvantaged counterparts under the pure rank order treatment. We also see large and significant gains in terms of time investment: disadvantaged subjects under treatment RQ increased investment by 170%. The other two measures capture specific tasks done during the time spent on the website: the number of topics attempted and the number of questions attempted. Although these two effects are less precisely estimated, both render large point estimates for increases of 99% each, and the latter is significant at the 10% level.

As for performance measures, we find a large and significant difference in final exam scores and exam score changes for disadvantaged subjects under RQ (i.e., preferential treatment): by both measures we estimate that they lifted their scores by roughly a fifth of a standard deviation, relative to their disadvantaged counterparts in the control group. Although some portion of this effect may be due to increased effort and concentration on the day of the final exam, we interpret this result and the other columns in Table 4 as evidence that treatment RQ induced additional learning through study.

One concern is that the strengthened incentives for group $\mathcal{D}$ subjects may come at the cost of weaker incentives for group $\mathcal{A}$ subjects, but this is not supported by the evidence in Table 4. For 4 out of 5 outcome measures, point estimates for the effect of RQ on group $\mathcal{A}$, given by $\beta_1 + \beta_2$, was actually positive, but insignificant. In column 5, the sum $\beta_1 + \beta_2$ 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 extensive margin of investment (an increase of 53.2%), with a

[12] One potential concern with including pre-test score is that it provides a noisy measure of initial human capital, and may, therefore, introduce an attenuation bias. As a robustness check we re-estimated the regressions in Table 4, omitting *Pre-Test*, and nearly all coefficient estimates and standard errors—except for the ones connected to β_3, the multiplier on higher-grade cohort status—remained virtually unchanged. See Table A.1 in the supplemental online appendix.

[13] Including school fixed effects accounts for some of the common factors shared by students in the same school. Since randomization occurred at the individual level and there were only 10 schools in our sample, we chose not to cluster our standard errors at the school level. Clustering at the classroom level would have been reasonable but we have classroom information for less than half our sample since the AMC8 was administered at the school level. Heteroskedasticity-robust standard errors are similar to those reported in Table 4, except that the p-value on *Total Time* rises to 0.110 while the p-value on # *Questions Attempted* drops to 0.083. Other heteroskedasticity-robust p-values are slightly smaller.

p-value of 0.103. Thus, the collective results above provide strong evidence in support of our theoretical prediction (**III**)—that group $\mathcal{D}$ increases average HC attainment under preferential treatment—and the data weakly favor an increase of investment for group $\mathcal{A}$ as well.[14]

4.2.3. ***Selective Attrition.*** One potential source of bias in our results concerning the performance measure (final exam score) is that 97 of the subjects who took the practice test and were randomly assigned to a competition group (9.8%) did not show up on the day of the final test.[15] Among disadvantaged subjects, those assigned to the *RQ* treatment were somewhat less likely to miss the final exam (10.6% vs. 16.7%). On the other hand, among disadvantaged subjects who did not show up for the final test, those assigned to the quota treatment had higher practice scores than those in the control *PRO* treatment (7.16 vs 6.35). However, practice scores among subjects who did show up for the final test were nearly the same across these two groups (7.91 vs 7.79). These comparisons all point in a direction opposite of our main results: if selective attrition is playing a role in our estimates, then the effect of the quota on final performance for disadvantaged subjects may have been greater.

4.2.4. ***Heterogeneous Treatment Effects by Ability Level.*** Recall that model prediction (**II**) is that for group $\mathcal{D}$ the test score distributions under *RQ* and *PRO* should have a unique interior crossing point, with the *RQ* CDF strictly above the *PRO* CDF to the right of the crossing point, and strictly below to the left of the crossing. In other words, there should be a positive mass of the most academically able subjects in group $\mathcal{D}$ who decrease output under a *RQ*, while subjects of middle and high learning costs increase output. Moreover, the upper bound on the output distribution should be lower for $\mathcal{D}$ under a *RQ*, with opposite predictions for group $\mathcal{A}$.[16] Note that since our within-group comparisons under *PRO* and *RQ* involve shifts in the intensity of competition while leaving other aspects of the contest fixed, our experiment implicitly tests model prediction (**I**) as well. Figure 1 depicts the comparison in plots of empirical CDFs for final exam scores by 7^{th} graders.[17] For the sake of comparability, these figures include only subjects for whom we have both pre-test and post-test scores.[18]

Point estimates of final exam score distributions in Figure 1 show patterns remarkably consistent with prediction (**II**) on quantile-specific effects within the disadvantaged group. In order to

[14]We also found no statistically significant differences by gender in test subjects' reactions to shifting competition.

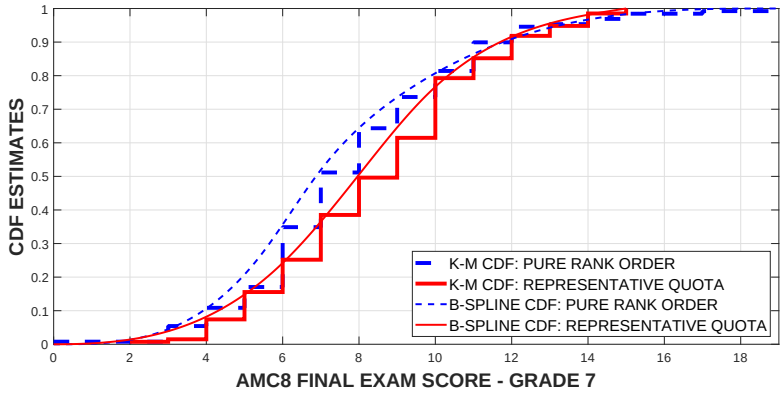
[15]Note: the effort measures did not require observing a final score. Table A.2 (online supplement) re-estimates the first 4 columns of Table 4 on the restricted sample. Treatment effects are very similar, but ∼10% noisier as one would expect.

[16]In this paper we explore distributions of both inputs and outputs, but it is important to remember that the predictions of the theory only directly apply to exam score (outputs) which are directly incentivized by the contest.

[17]For a baseline comparison, Figure A.1 (online supplement) plots pre-test CDFs for 7^{th} and 8^{th} graders. Group $\mathcal{D}$ on average had to achieve more progress in order to receive a prize. This is not the same as observing costs, but the two are related and the hypothesis of stochastic dominance in cost types which underpins the model predictions appears plausible.

[18]Figures A.3 and A.4 in the Online Appendix contain plots comparing pre-test scores for all test subjects. They suggest that selective attrition is generally working against our results presented here. For group $\mathcal{D}$ the pre-test *PRO* distribution is slightly below the pre-test *RQ* distribution for values at or below the median, and the upper bound of the pre-test distribution for 7^{th} grade *RQ* subjects is higher. Both of these patterns are substantially reversed by the final exam.

FIGURE 1. SEVENTH GRADE FINAL EXAM SCORES



Notes: For the sake of comparability, Figure 1 uses only data for 7th graders who took both the pre-test *and* final exam. See Figure A.3 in the online appendix for a discussion on the role of selective attrition. Figure A.2 in the online appendix contains an analogous plot for 8th grade final exam scores by treatment group.

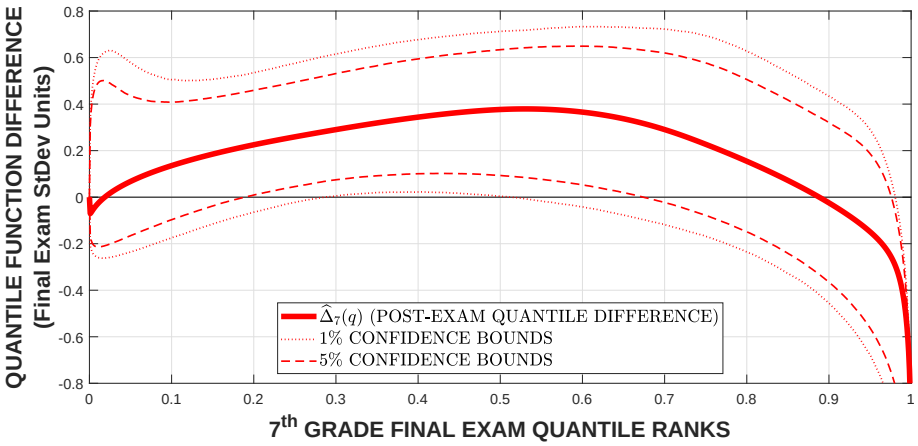
formally investigate the strength of the evidence we wish to make within-group, cross-treatment comparisons which allow for the magnitude and sign of the effect to vary across different quantiles. For this purpose we estimate smoothed quantile functions $\widehat{\mathcal{H}}_j^t(q)$ based on flexible cubic B-spline CDFs for each cohort group $j = \mathcal{A}, \mathcal{D}$ and treatment $t = PRO,\ RQ$.[19] Figure 1 displays a comparison of the Kaplan-Meier empirical CDFs (thick lines) and the smoothed, B-spline CDFs (thin lines) for 7^{th}-graders. With the implied quantile functions we can compute a difference, $\widehat{\Delta}_j(q) = \widehat{\mathcal{H}}_j^{RQ}(q) - \widehat{\mathcal{H}}_j^{PRO}(q)$, which represents the final exam score difference at the q^{th} quantile of group $j = \mathcal{A}, \mathcal{D}$ under a RQ rule (i.e., affirmative action for 7th graders), relative to baseline PRO competition.

In order to account for sampling variability we execute a bootstrap routine, wherein we re-sampled separately from each age-group/treatment subsample (with replacement) and re-estimated $\widehat{\Delta}_j(\cdot)$ 50,000 times. Figure 2 displays the point-estimate quantile difference with 95% confidence bounds (thin dashed lines) and 99% confidence bounds (thin dotted lines). Quantile ranks $q \in [0, 1]$ are on the horizontal axis, and differences are on the vertical axis in final exam standard deviation units. Point estimates for group $\mathcal{D}$ conform well to model predictions (**I**) and (**II**): we see an interior point near the top of the exam score distribution (89th percentile) where $\widehat{\Delta}_{\mathcal{D}}(\cdot)$ crosses the zero line. Moreover, achievement under RQ within the inter-quartile range improves by 20% – 40% of a standard deviation, while above the 95th percentile we get reductions of 20% – 40% of a standard deviation.

The plotted confidence bounds illustrate the strength of the statistical evidence for the quantile-specific differences. At the 95% level, the predicted achievement increase for medium and lower

[19]The central tuning parameters that determine mean integrated squared error of B-Spline CDFs are the number and placement of knots. Following best practice, knots were uniformly spaced in quantile ranks of the raw data, and we chose a 5-knot version of our cubic B-Spline estimator (implying 6 estimated free parameters), as this minimized both the Akaike and Bayesian information criteria. See the Online Supplemental Appendix for additional details.

FIGURE 2. 7th-GRADE FINAL EXAM QUANTILE DIFFERENCES



ability students is statistically different from zero between the 20th and 70th percentiles, and the predicted reduction in achievement by the best and brightest students is significant above the 97th percentile. As a supplemental test of the evidence for prediction (**II**), we also executed a parametric bootstrap test for the null hypothesis that the top student in the $(\mathcal{D}, RQ)$ group weakly outscores the top $(\mathcal{D}, PRO)$ student, against the one-sided alternative that prediction (**II**) holds at the upper bound instead (i.e., the top student performs best under PRO). We reject the null hypothesis with a p-value of 0.0619, providing further evidence in favor of the heterogeneous treatment effects characterized by predictions (**I**) and (**II**).[20] Quantile difference point estimates in group $\mathcal{A}$ are also consistent with theory though smaller and statistically insignificant (see Figure A.2 in the online supplemental appendix).

4.2.5. ***Narrowing Achievement Gaps***. We now look at the tendency for a preferential treatment policy to narrow achievement gaps across demographic groups. Table 5 displays summary statistics on 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 all grade 7 and 8 subjects.[21] Therefore, the means indicate distance between the population average and grade cohort average, in standard deviation units. Without accounting for treatment status, 7th grade subjects were roughly half of a standard deviation behind their 8th grade counterparts—or $-0.295 - 0.181 = -0.476$ standard deviations to be exact—but by the final exam, the gap between 7th and 8th graders had narrowed by about half—to $-0.138 - 0.085 = -0.223$ standard deviations.

[20] We use a parametric bootstrap test because the granularity of the raw data becomes a hindrance when zooming in on a narrow segment of the distribution. The test was based on the actual age cohort/treatment sample sizes, and our optimized cubic B-Spline CDFs. See Supplemental Online Appendix B.2 for full details and discussion on our parametric bootstrap test procedure. An alternate parametric bootstrap test based on boundary-corrected, kernel-smoothed CDFs (to ensure reliable tail behavior; see Karunamuni and Zhang (2008) and Hickman and Hubbard (2015)) produces a slightly stronger result, rejecting the null with a p-value of 0.0332 instead.

[21] In order to make the pre-test and final exam figures comparable, we excluded subjects whose final scores were missing.

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	0.070	0.960	431
Standardized Final Score (GRADE 7)	-0.138	-0.205	1.005	264
Standardized Final Score (GRADE 8)	0.085	0.142	0.989	431
Achievement Gaps for Representative Quota Treatment				
Standardized Pre-Score (GRADE 7)	-0.275	-0.262	1.063	135
Standardized Pre-Score (GRADE 8)	0.169	0.073	0.922	220
Standardized Final Score (GRADE 7)	-0.052	0.117	0.945	135
Standardized Final Score (GRADE 8)	0.032	0.117	1.033	220
Achievement Gaps for Pure Rank Order (Control) Treatment				
Standardized Pre-Score (GRADE 7)	-0.317	-0.272	0.920	129
Standardized Pre-Score (GRADE 8)	0.194	0.068	0.999	211
Standardized Final Score (GRADE 7)	-0.231	-0.537	1.059	129
Standardized Final Score (GRADE 8)	0.141	0.168	0.937	211

Notes: Each panel in the table contains standardized scores on the pre-test and post-test, where standardization was performed within each panel-test grouping, excluding scores for subjects who missed the final exam. The table includes only information on 7^{th} and 8^{th} graders for comparability since those two groups competed head-to-head in the *PRO* treatment. Moreover, 5^{th} and 6^{th} graders in our study came from accelerated math classes (and make up only 20% of our sample), while 7^{th} and 8^{th} graders came from regular math classes.

In the lower two panels of Table 5 we break out this effect by treatment group, but with score standardization now happening within each exam-treatment cell. Part of the test score convergence had to do with differences in the conditions of the pre-test and final exams (likely due to incentives or slightly different content): within the *PRO* treatment, about a quarter of the gap disappeared but remained at 0.372 final exam standard deviations. However, the achievement gap under the *RQ* closed substantially more, by about 80%, beginning at 0.444 standard deviations, and ending at only 0.084 standard deviations on the final exam.[22] Not only are the changes in achievement gaps large, but they are statistically significant as well. Within the RQ treatment (middle panel in Table 5) the difference in means for the pre-test is significant at the 1% level (t-statistic=4.15), while the mean difference in the final exam score is not (t-statistic=0.77). In contrast, for the PRO treatment (lower panel of Table 5), the difference in exam performance is statistically significant both on the pre-test (t-statistic=4.71) and on the final exam (t-statistic=3.38). These findings further support theoretical predictions (**II**) and (**III**) which together imply that preferential treatment in HC investment contests can help to narrow achievement gaps through improved investment incentives for a disadvantaged target demographic.

[22]The within-treatment-demographic standard deviations are all close to one, which suggests the narrowing of gaps within the two treatments was due predominantly to mean/median shifts in test scores.

5. DISCUSSION AND CONCLUSION

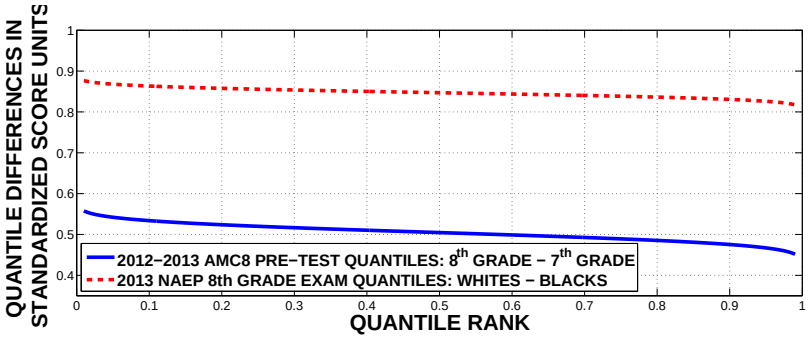
5.1. **Real-world interpretations and limitations.** Our paper has focused on AA in college admissions, and how such policies change students' incentives to invest in productive HC before applying to college. However, the experimental design and results also provide insight into how investment behavior changes in response to more general shifts in the intensity of competition for high-quality match partners. This latter phenomenon arises in many real-world applications, including college admissions and labor markets, even in the absence of AA. For example, it has been widely speculated that today's admissions process for elite American universities is an increasingly burdensome academic arms race that imposes excessive and unnecessary costs on students and their families. Our experimental results show that students who were already the high achievers under the status quo further increase investment effort when competition becomes more intense, even if the intrinsic costs and benefits of HC do not change. Examples that might lead to the increasing intensity of competition include the proliferation of new investment technologies—*e.g.*, advanced-placement courses and extra-curricular academic programs—and universities marketing their services to top students from abroad to improve revenues, which would tend to select additional low-cost agents into the competition.

The purpose of this paper is not to make claims about specific magnitudes of these effects in any particular context. Rather, we provide evidence for the empirical relevance of relative incentives for shaping investment behavior, an idea which has been largely overlooked in the human capital literature. A related question is whether agents display the sophistication needed to react to subtle changes in their outside competition. As we have seen, even middle-school children appear capable of behavioral patterns that conform remarkably well to Bayes-Nash equilibrium play under shifting relative incentives.

5.1.1. *External Validity: Sample Population.* When we discuss the insights of our results regarding AA specifically, it is important to consider what our analysis can and cannot say. The ultimate question of interest is the effect that real-world AA has on academic incentives in practice. There are several reasons why this study cannot fully address this question. First, our partner schools for this study serve a fairly homogeneous population of students. On one hand, defining our disadvantaged/advantaged demographic group as lower-/higher-grade students ensures that we can easily understand the differences between them: advantaged group students are just like their counterparts, except they are a year older and have received one extra year of schooling. This allows for a clean exploration of theoretical insights. On the other hand, we are clearly missing some factors such as significant variation in cultural norms or school quality that could affect how behavior changes in settings other than suburban Utah. Various social phenomena contribute to the complexity of real-world AA environments where disadvantaged demographic groups face challenges or barriers that may not be present for advantaged groups.[23]

[23]E.g., Cotton et al. (2020) Find strong evidence of a causal link between school quality differences and racial achievement gaps among suburban fifth and sixth graders in Illinois.

FIGURE 3. GRADE-LEVEL DIFFERENCES VS BLACK/WHITE DIFFERENCES



NOTES: This figure first standardizes AMC8 and NAEP exam scores by subtracting the mean and dividing by the standard deviation for the relevant sample population. Then, the distribution of standardized test scores for each sub-sample–*i.e.*, AMC8 7^{th} grade, AMC8 8^{th} grade, NAEP 8^{th} grade whites, and NAEP 8^{th} grade blacks–is approximated by a normal distribution. The resulting quantile functions are subtracted from each other in order to illustrate relative, quantile-specific disparities between groups in standard deviation units.

Figure 3 provides a comparison between our sample pool and the broader US population of pre-college HC investors using data from the NAEP (National Assessment of Educational Progress) exam. The graph compares the difference between black and white 8^{th} grade NAEP quantiles—expressed in standardized NAEP score units (dashed line)—to the difference in 7^{th} and 8^{th} grade quantiles from our sample—expressed in standardized AMC8 score units (solid line).[24] At most NAEP quantiles, black 8^{th} graders trail their white counterparts by about 0.85 standard deviations, whereas the gap among our test subjects is closer to 0.5 standard deviations. Thus, our grade-level delimiter between the advantaged and disadvantaged groups likely understates the competitive difference between blacks and whites vying for actual college seats, and therefore real-world discouragement effects minorities face could be larger than those found in our experimental study. However, other factors may play a role, including differences in home environments, gender norms, cultural factors, or behavioral phenomena (e.g., stereotype threat) which may push the results in other, less predictable directions. Certainly though, in order to understand how and why a policy works, it is important to cleanly understand the incentive dimension in addition to other sociocultural factors.

5.1.2. *External Validity: Short-Run vs. Long-Run Incentives.* A second limitation in the current analysis is that, although it creates incentives which mirror HC competition in some key ways, these are only engineered on a small scale and measured over a short-run horizon. We tracked students over a 10-day period during which they had the opportunity to invest leisure time into their accumulated math proficiency. For students in our sample who logged positive amounts of study time, their average expected wage was $10.73/hour (=total earnings/total study time).

[24]The NAEP is administered to a random sample of all 8^{th} graders nationally (sampled at the classroom level), and therefore includes many students who will not ultimately apply for college. However, this actually fosters comparisons to our sample population: we recruited regular 7^{th} and 8^{th} grade math classrooms for our study, with participation on an opt-out basis, so our sample also spanned the college-going and non-college-going segments of the Utah population.

Students preparing for college optimize labor-leisure division over a much longer horizon (4-5 years until college), with much larger payoffs that come far in the future.

In section B.5 of the Online Appendix, we calibrate some rough measures of possible time use incentives for US high school students under varying assumptions on time discounting, study intensity, college graduation rates, and the cost of college. This rough calibration results in an effective hourly compensation rate of between \$10 and \$55 per hour of study for an average high-school student. Although somewhat crude, these calculations suggest that the short-run incentives in our experimental study may not be entirely incongruous with the magnitudes of incentives actual high-school students face to study for college. Quantifying pre-college incentives over years of childhood is complex, but this study underscores the relevance of competitive incentives in investment decisions for productive HC.

5.2. **Conclusion.** To explore how students respond to AA incentives, we executed a field experiment in which hundreds of students compete in mathematics achievement for heterogeneous cash prizes, while allowing students a 10-day period in which they can study and prepare for the contest. Our experimental design includes a preferential *RQ* treatment in favor of systemically disadvantaged students, with time-use monitoring to measure how competitive incentives shape student investment in developing HC. We chose a representative quota for this study because of its relative simplicity, though the space of all AA mechanisms is vast and covers many possibilities. However, our experimental evidence serves as a proof of concept that preferential treatment can be used as a means of positive incentive provision for disadvantaged demographic groups.

Traditional wisdom has held that AA (in any form) will generally erode incentives for the group it seeks to aid by lowering admissions standards. Our results offer a rebuttal to this idea: a well-designed preferential treatment policy which targets disadvantaged students may actually *increase* average effort by placing within reach contest outcomes which would otherwise be unattainable. This mitigates discouragement effects that arise when disadvantaged agents find themselves too far behind the competitive curve. Moreover, conventional wisdom in the labor economics literature has held that resource asymmetry creates achievement gaps in a mechanical way, simply by creating gaps in HC production technology. However, the comparative static scenario in our work effectively holds individual production technology fixed, but still produces meaningful changes in learning behaviors by altering the distribution of one's competitors. In short, our study suggests that academic achievement gaps between students with asymmetric backgrounds are endogenous equilibrium objects, and should not be thought of as exogenous or mechanical.

References

Akaike, Hirotugu. 1973. "Information Theory as an Extension of the Maximum Likelihood Principle." *Second International Symposium on Information Theory*, , ed. B.N. Petrov and F. Csaki, 267–281.

Akhtari, Mitra, Natalie Bau, and Jean-William Laliberté. 2020. "Affirmative Actio nand Pre-College Human Capital." *working paper, UCLA Economics Department*.

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

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

Bodoh-Creed, Aaron, and Brent R. Hickman. 2018. "College Assignment as a Large Contest." *Journal of Economic Theory*, 175: 88–126.

Bodoh-Creed, Aaron, and Brent R Hickman. 2019. "Identifying the Returns to College Quality Using Affirmative Action." *working paper, Washington University in St. Louis, Olin Business School*.

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

Burgess, Simon, Robert Metcalfe, and Sally Sadoff. 2016. "Understanding the Response to Financial and Non-Financial Incentives in Education: Field-Experimental Evidence Using High-Stakes Assessments." IZA Discussion Paper Series.

Burnham, K., and D. Anderson, ed. 2002. *Model Selection and Multimodal Inference.* New York: Springer.

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

Chambers, David L., Timothy T. Clydesdale, William C. Kidder, and Richard O. Lempert. 2005. "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.

Coate, Stephen, and Glenn Loury. 1993. "Anti-Discrimination Enforcement and the Problem of Patronization." *American Economic Review*, 83(2): 92–98.

Cotton, Christopher S., Brent R. Hickman, and Joseph P. Price. 2020. "Affirmative Action, Shifting Competition, and Human Capital Accumulation: A Comparative Static Analysis of Investment Contests." *Queen's University working paper*.

Cotton, Christopher S., Brent R. Hickman, John A. List, Joseph P. Price, and Sutanuka Roy. 2020. "Productivity Versus Motivation in Adolescent Human Capital Production: Evidence from a Structurally-Motivated Field Experiment." *NBER Working Paper 27995*.

de Boor, Carl, ed. 2001. *A Practical Guide to B-Splines, Revised Edition.* New York: Springer-Verlag.

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

Hickman, Brent R., and Timothy P. Hubbard. 2015. "Replacing Sample Trimming with Boundary Correction in Nonparametric Estimation of FirstPrice Auctions." *Journal of Applied Econometrics*, 30(5): 739–762.

Hickman, Brent R., Timothy P. Hubbard, and Harry J. Paarsch. 2017. "Identification and Estimation of a Bidding Model for Electronic Auctions." *Quantitative Economics*, 8(2): 505–551.

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

Karunamuni, Rhoana J., and Shunpu Zhang. 2008. "Some Improvements on a Boundary Corrected Kernel Density Estimator." *Statistics & Probability Letters*, 78: 499–507.

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

Leuven, Edwin, Hessel Oosterbeek, and Bas van der Klaauw. 2010. "The Effect of Financial

Rewards on Students' Achievement: Evidence from a Randomized Experiment." *Journal of the European Economic Association*, 8: 1243–1265.

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

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

Moro, Andrea, and Peter Norman. 2003. "Affirmative Action in a Competitive Economy." *Journal of Public Economics*, 87: 567–594.

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

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

Schwarz, Gideon E. 1978. "Estimating the Dimension of a Model." *Annals of Statistics*, 6(2): 461–464.

Appendix A. SUPPLEMENTAL ONLINE APPENDIX

To Accompany *Affirmative Action and Huma Capital Invesment: Evidece from a Randomized Field Experiment*,
by Christopher S. Cotton, Brent R. Hickman, and Joseph P. Price

A.1. **ROBUSTNESS CHECKS.**

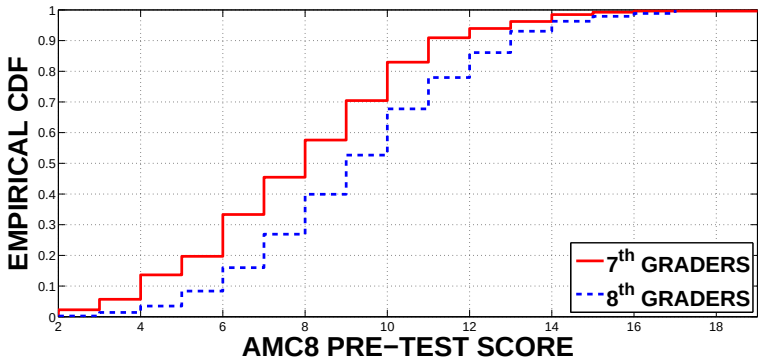
A.1.1. *Testing Average Differences.* Table 4 tests for average treatment differences using standardized pre-test score as a control. One potential concern with including pre-test scores in the regressions for Table 4 is that it provides a noisy measure of initial human capital, and may therefore introduce an attenuation bias. As a robustness check we re-estimated the regressions in Table 4, omitting pre-test as a control, and nearly all coefficient estimates and standard errors—except for the ones connected to β_3, the multiplier on higher-grade cohort status—remained virtually unchanged. See Table A.1 displays the results of this additional analysis. Table A.1 below re-estimates the regressions in Table 4, omitting pre-test score as a control (see comment in footnote 12 in the body).

Table A.1. (RE-)TESTING DIFFS BY DEMOGRAPHICS AND TREATMENT

	Investment				**Performance**	
	Used Website	# Subjects Attempted	Total Time	# Questions Attempted	Final Exam Score	Exam Score Change
Constant $(\widehat{\beta}_0)$	0.069***	0.138**	3.009	1.157	7.750***	0.096
Std. Err.	(0.024)	(0.059)	(1.951)	(0.730)	(0.229)	(0.239)
Quota $(\widehat{\beta}_1)$	0.091***	0.154*	5.697**	1.373	0.646**	0.591*
Std. Err.	(0.033)	(0.083)	(2.754)	(1.030)	(0.318)	(0.330)
P-Value:	*[0.006]*	*[0.064]*	*[0.039]*	*[0.183]*	*[0.042]*	*[0.074]*
Advantaged $*$ *Quota* $(\widehat{\beta}_2)$	-0.050	-0.117	-5.197	-0.928	-0.862**	-0.546
Std. Err.	(0.042)	(0.105)	(3.503)	(1.311)	(0.399)	(0.415)
Advantaged $(\widehat{\beta}_3)$	0.034	0.090	1.621	1.128	1.295***	-0.324
Std. Err.	(0.030)	(0.074)	(2.473)	(0.926)	(0.284)	0.300
School Fixed Effects	yes	yes	yes	yes	yes	yes
N	992	992	992	992	895	895
Additional Test: Effect of Quota on Advantaged Group						
$\widehat{\beta}_1 + \widehat{\beta}_2$	0.041	0.037	0.500	0.445	-0.216	0.045
P-Value:	*[0.118]*	*[0.570]*	*[0.818]*	*[0.583]*	*[0.371]*	*[0.857]*

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). Standard errors are in parentheses. 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.

FIGURE A.1. PRE-TEST SCORES: 7^{th} GRADE VS 8^{th} GRADE



Notes: This figure only includes data from students who completed both the pre-test and the post-test (see next section below for an exploration of selective attrition).

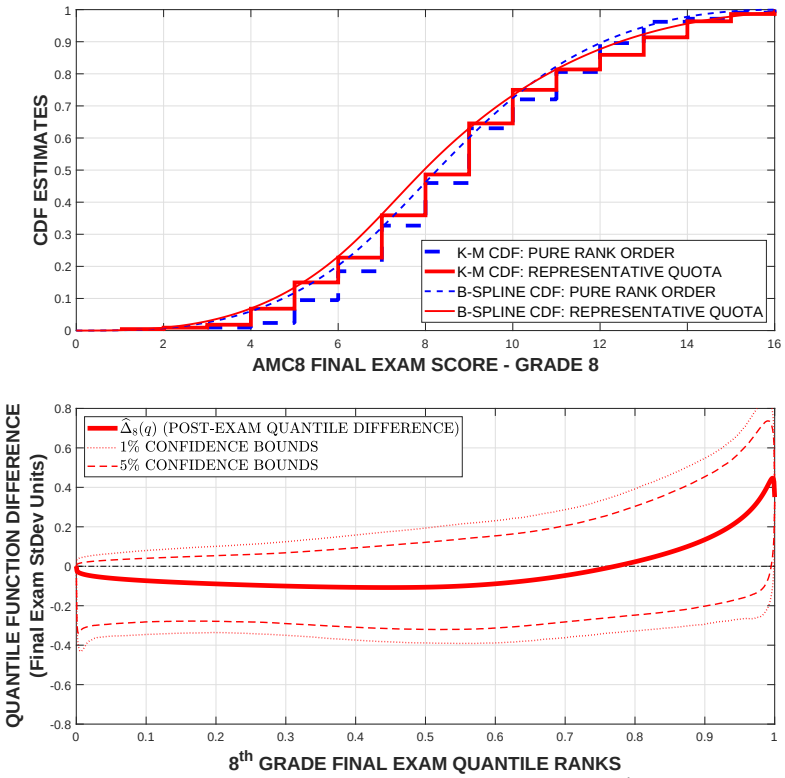
A.1.2. *Pre-Test Scores.* Figure A.1 plots the CDFs of pre-test scores for grades 7 and 8. The figure strongly supports stochastic dominance of initial math proficiency levels across demographic groups. A Kolmogorov-Smirnov test rejects the null hypothesis that the 7^{th} and 8^{th} grade distributions are the same, against a one-sided alternative of stochastic dominance with a p-value of 1.03×10^{-5}. This means group $\mathcal{D}$ subjects, on average, had to achieve more progress in order to be competitive for a prize. This is not the same as observing costs, but the two are certainly related and the hypothesis of stochastic dominance in cost types appears plausible.

A.1.3. 8^{th}*-Grade Quantile-Specific Changes.* Figure A.2 plots the CDFs of post-test scores for grade 8 test subjects under a PRO regime and under a RQ regime. The point estimate quantile difference function conforms to prediction (**II**) from the theory model, though the measured shifts are small and the confidence bounds indicate that they are noisy as well.

A.1.4. *Selective Attrition.* Figures A.3 – A.4 illustrate a robustness check on our quantile function estimator, when we attempt to adjust for selective attrition. The upper panels in Figures A.3 and A.4 plot the empirical CDFs of pre-test scores, restricted to the subsample of students who took the final exam as well. The bottom panels re-produce the CDFs of final exam scores for comparison. From these figures it appears that selective attrition may be working slightly against finding our results in general.

Figures A.5 and A.6 are an attempt at adjusting our quantile function estimator for the possible influence of selective attrition. To do so, we begin by computing the quantile difference function for post-test scores, $\widehat{\Delta}_j(q),\ j = \mathcal{A}, \mathcal{D}$ measured in exam standard deviation units, restricted only to the sample of test subjects who took the final exam as explained in the body. Then, and we compute two alternate versions of the quantile difference function for pre-test scores. The first, which we shall call $\widehat{\Delta}_j^{pre,partial}(q),\ j = \mathcal{A}, \mathcal{D}$, is the quantile difference function (across treatments, but within age cohort) for pre-test scores (expressed in standard deviation units) using

FIGURE A.2. EIGHTH GRADE FINAL EXAM SCORES



Notes: For the sake of comparability, Figure A.2 uses only data for 8th graders who took both the pre-test *and* final exam. See Figure A.4 for a discussion on the role of selective attrition.

FIGURE A.3. 7th GRADE BY TREATMENT:
Pre-Test vs. Final Exam

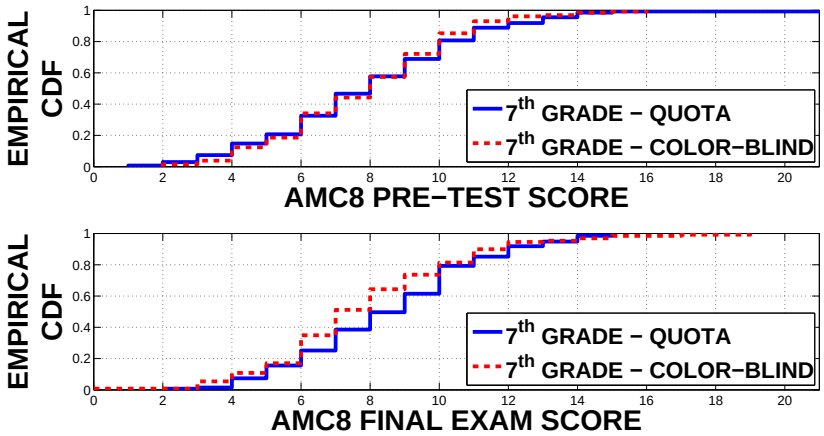


FIGURE A.4. 8^{th} GRADE BY TREATMENT:
Pre-Test vs. Final Exam

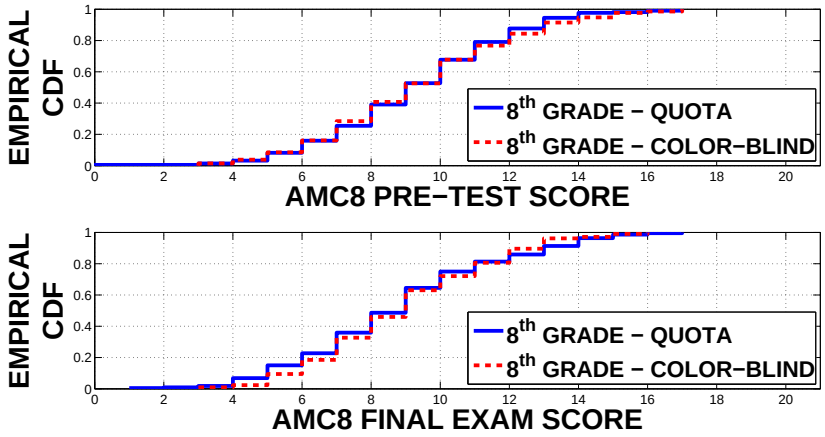
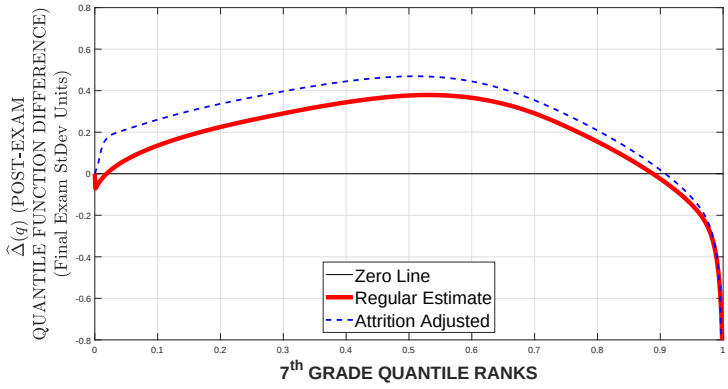


FIGURE A.5. 7^{th} GRADE QUANTILE DIFFERENCES, ADJUSTED FOR ATTRITION

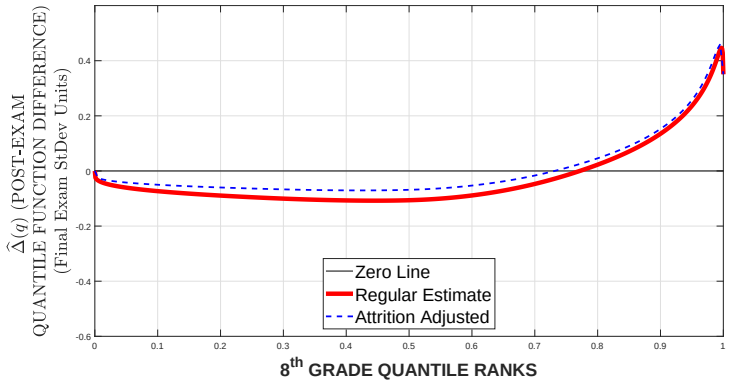


the partial sample of subjects who showed up for the final exam. The second, which we shall call $\widehat{\Delta}_j^{pre,full}(q)$, $j = \mathcal{A}, \mathcal{D}$, is the quantile difference function for pre-test scores (expressed in standard deviation units), but using the full sample of all students who took the pre-test and were randomized into a treatment cell. We then adjust the post-test quantile difference function for attrition by subtracting out the pre-test difference in the quantile difference function across the non-attriter partial sample and the full sample:

$$\widehat{\Delta}_j^{adjusted}(q) = \widehat{\Delta}_j(q) - \left(\widehat{\Delta}_j^{pre,partial}(q) - \widehat{\Delta}_j^{pre,full}(q)\right).$$

Figure A.5 shows a comparison between the attrition-adjusted quantile difference function and the regular estimate discussed in the body of the paper. The former point estimate shows a pattern that is slightly stronger than the regular estimate we report in the body. On the other hand, it is also a noisier estimate since it is a sum of three separately estimated quantile difference functions rather than only one.

FIGURE A.6. 8^{th} GRADE QUANTILE DIFFERENCES, ADJUSTED FOR ATTRITION



As another check regarding the possible role of selective attrition, recall that Table 4 in the body of the paper contains 6 columns. The first 4 columns are for effort measures and use the full sample ($N = 992$), while the last 2 columns are for outcome measures (final exam score and pre-post score change), and therefore use only the subsample for which final exam scores are available ($N = 895$). As a final robustness check, we re-estimate the regressions in the first 4 columns below, using the restricted sample of non-attriters, which is roughly 10% smaller than the full sample. These results are reported in Table A.2 in this section. The point estimates of treatment effects (β_1) are nearly identical to those reported in Table 4, while the standard errors on treatment effects are roughly 10% higher. This change is roughly in line with what one would expect if the dropped observations were merely reducing sample size and not inducing selection bias.

APPENDIX B. EMPIRICAL ANALYSIS: ADDITIONAL TECHNICAL DETAILS, FIGURES, AND TABLES

B.1. **B-SPLINE CDF ESTIMATOR DETAILS.** In order to get an estimate of the distributions of AMC8 scores that allows for comparisons at arbitrary quantiles across treatments groups, we smooth them using a B-spline representation of the AMC8 exam score CDFs. B-splines are a parametric family of functions which combines remarkable flexibility with numerical stability. They are mathematically equivalent to piece-wise, local polynomials (i.e., regular splines), but afford greater numerical convenience, being made up of globally-defined basis functions, like orthogonal polynomials (e.g., Chebyshev), but without exhibiting their often erratic behavior. Moreover, incorporating shape restrictions, such as monotonicity and terminal conditions, is quite simple within the B-spline family.[25]

The first step is to define a *knot vector* which partitions the support into sub-intervals over which a local cubic polynomial will represent the shape of the CDF. Specifically, within each treatment $t \in \{RQ, PRO\}$ and group $j \in \{\mathcal{A}, \mathcal{D}\}$ we partition the support of final exam scores $[0, \overline{h}_j^t]$ into K sub-intervals and we specify the breakpoints between these sub-intervals as a vector

[25] A standard text on B-splines is de Boor (2001). For a brief primer on B-splines and their advantages for empirical economics, see Hickman, Hubbard and Paarsch (2017, Appendix).

TABLE A.2. (RE-)TESTING EFFORT DIFFS BY DEMOGRAPHICS AND TREATMENT ON THE RESTRICTED SAMPLE OF NON-ATTRITERS

	Investment			
	Used Website	# Subjects Attempted	Total Time	# Questions Attempted
Constant ($\widehat{\beta}_0$)	0.086***	0.173***	3.741*	1.555*
Std. Err.	(0.027)	(0.066)	(2.231)	(0.829)
Quota ($\widehat{\beta}_1$)	0.086**	0.147	5.987*	1.236
Std. Err.	(0.037)	(0.091)	(3.084)	(1.145)
P-Value:	*[0.019]*	*[0.109]*	*[0.053]*	*[0.281]*
Advantaged $*$ *Quota* ($\widehat{\beta}_2$)	-0.032	-0.083	-4.957	-0.541
Std. Err.	(0.046)	(0.115)	(3.872)	(1.438)
Advantaged ($\widehat{\beta}_3$)	0.010	0.042	0.602	0.557
Std. Err.	(0.033)	(0.083)	(2.809)	(1.043)
$Pre-Test$ ($\widehat{\beta}_4$)	0.028**	0.036	0.755	0.457
Std. Err.	(0.012)	(0.029)	(0.984)	(0.365)
School Fixed Effects	yes	yes	yes	yes
N	895	895	895	895

Notes: Each column is a separate regression. Advantaged is an indicator variable for whether the student is a 6th or 8th grader (the older group in each school type). Standard errors are in parentheses. 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.

of $K+1$ *knots* (including endpoints). The knot vector, $\mathbf{k}_j^t = \{k_{j,1}^t = 0, k_{j,2}^t, \ldots, k_{j,K}^t, k_{j,K+1}^t = \overline{h}_j^t\}$ (specified in non-decreasing order), uniquely defines a set of $K+3$ cubic B-spline basis functions through the Cox-de Boor recursion formula (see de Boor (2001)). The resulting basis functions, denoted $\mathcal{B}_{j,k}^t : [0, \overline{h}_j^t] \to \mathbb{R}$, $k = 1, \ldots, K+3$, form the basic building blocks of our flexible parametric form for the CDF of exam scores:

$$G_j^t\left(h; \boldsymbol{\alpha}_j^t\right) = \sum_{k=1}^{K+3} \alpha_{j,k}^t \mathcal{B}_{j,k}^t(h).$$

We estimate the parameter vector $\boldsymbol{\alpha}_j^t$ via a GMM routine which achieves a constrained, least-squares, best fit to the empirical CDF,

$$\begin{aligned} \widehat{\boldsymbol{\alpha}}_j^t \equiv \min_{\boldsymbol{\alpha}_j^t} & \left\{ \sum_{n=1}^{N_j^t} \left(\widehat{G}_j^t(h_{j,n}^t) - G_j^t\left(h_{j,n}^t; \boldsymbol{\alpha}_j^t\right) \right)^2 \right\} \\ & \textit{subject to}: \\ & \alpha_{j,1}^t = 0, \;\; \alpha_{j,K+3}^t = 1 \\ & \alpha_{j,k}^t < \alpha_{j,k+1}^t, \;\; k = 1, \ldots, K+2. \end{aligned} \tag{B.1}$$

In the objective function above, $\widehat{G}_j^t(h_{j,n}^t) = \sum_{l=1}^{N_j^t} \mathbf{1}(h_{j,l}^t \leq h_{j,n}^t)/N_j^t$ is the Kaplan-Meier empirical CDF of AMC8 scores for group $j = \mathcal{A},\ \mathcal{D}$ under treatment $t = RQ,\ PRO$, evaluated at each datapoint $h_{j,n}^t$ in the relevant subsample. The first two constraints enforce terminal contitions: $G_j^t(0;\boldsymbol{\alpha}_j^t) = 0$ and $G_j^t\left(\overline{h}_j^t;\boldsymbol{\alpha}_j^t\right) = 1$. The second line of the constraints ensures strict monotonicity of the CDF.

The attractive property of B-Splines is not just their numerical stability and computational convenience, but the fact that they can be used to achieve arbitrary degrees of flexibility by adding additional knots to the knot vector. This produces a finer partition of the support, more basis functions, and thus more flexibility of shapes that can feasibly be recovered by the B-Spline-based estimator. In that sense, our method belongs to the class of *sieve estimators* since, if we allow the number of knots $(K+1)$ to grow with the sample size N_j^t (in such a way that $\mathbf{k}_j^t$ becomes a dense set in the limit as $N_j^t \longrightarrow \infty$), a B-spline CDF can eventually accommodate arbitrary shapes.

In finite samples, the crucial tuning parameters which determine the statistical properties of a B-Spline CDF estimator are the number and placement of the knots. As for knot placement, for a given K we select $K+1$ knots at uniformly-spaced locations in quantile ranks, or

$$k_{j,l}^t = \widehat{\mathcal{H}}_j^t\left(\frac{l}{K}\right),\ \ l = 0, 1, 2, \ldots, K,$$

where $\widehat{\mathcal{H}}_j^t$ is a linear interpolation of the empirical quantile function evaluated at the sample data points. Uniform spacing in quantile ranks ensures that the influence of the data are uniformly spread over the various parameters to be estimated, which is illustrated in Figure B.1 below. Intuitively, under this rule if K is chosen to be 4, then the knots are placed at the empirical quartiles, whereas if $K = 10$ we get knots at the empirical deciles, and so on.

The only remaining choice is how many sub-intervals K, which involves a familiar statistical trade-off: the higher is K, the more flexible the B-Spline CDF and therefore the lower the bias, but with higher K also comes increased variance as well. To resolve this issue, we use some well-known tools for model selection: the Akaike information criterion (Akaike (1973)),

$$AIC_{j,K} = 2(2K+2) - 2\widehat{\mathcal{L}}_{j,K}^{PRO} - 2\widehat{\mathcal{L}}_{j,K}^{RQ},$$

and the Bayesian information criterion (Schwarz (1978)),

$$BIC_{j,K} = \ln\left(N_j^{PRO}\right)(K+1) + \ln\left(N_j^{RQ}\right)(K+1) - 2\widehat{\mathcal{L}}_{j,K}^{PRO} - 2\widehat{\mathcal{L}}_{j,K}^{RQ},$$

where $\widehat{\mathcal{L}}_{j,K}^t$ represents the log-likelihood function evaluated at the estimated parameters $\widehat{\boldsymbol{\alpha}}_{j,K}^t$, and $(K+1)$ represents the number of free parameters to be estimated for each age-treatment pair $(t,j) \in \{PRO, RQ\} \times \{\mathcal{A}, \mathcal{D}\}$, once the terminal conditions are imposed.[26] For various candidate

[26]Note that the AIC and BIC formulae above apply when choosing K to simultaneously optimize the statistical properties of $G_j^{PRO}(h;\widehat{\boldsymbol{\alpha}}_j^{PRO})$ and $G_j^{RQ}(h;\widehat{\boldsymbol{\alpha}}_j^{RQ})$ together. The reason why K is chosen to optimize both CDFs together is because they each form part of the quantile difference function $\widehat{\Delta}_j(q)$, the object of primary interest.

TABLE B.1. B-SPLINE CDF MODEL SELECTION (7^{th} GRADE)

MODEL	# OF PARAMETERS/ BASIS FUNCTIONS	$BIC_{7,K}$	$AIC_{7,K}$	$\Pr[K \text{ optimal}\|AIC_{7,K}]$
$K = 2$	5	1,465.08	1,447.81	$< 10^{-10}$
$K = 3$	6	1,373.62	1,350.59	0.00382263
$\mathbf{K^* = 4}$	**7**	**1,368.24**	**1,339.46**	**1.00000000**
$K = 5$	8	1,377.47	1,342.92	0.17649827
$K = 6$	9	1,380.37	1,340.06	0.73855115
$K = 7$	10	1,414.43	1,368.37	0.00000053
$K = 8$	11	1,413.58	1,361.65	0.00001433
$K = 9$	12	1,433.48	1,375.90	0.00000001

values of K, the one that minimizes AIC and/or BIC indicates the model which is approximately most likely to minimize mean integrated squared error (MISE) of the B-Spline CDF, relative to the truth. In the case of $AIC_{j,K}$, we get a convenient formula (due to Burnham and Anderson (2002)),

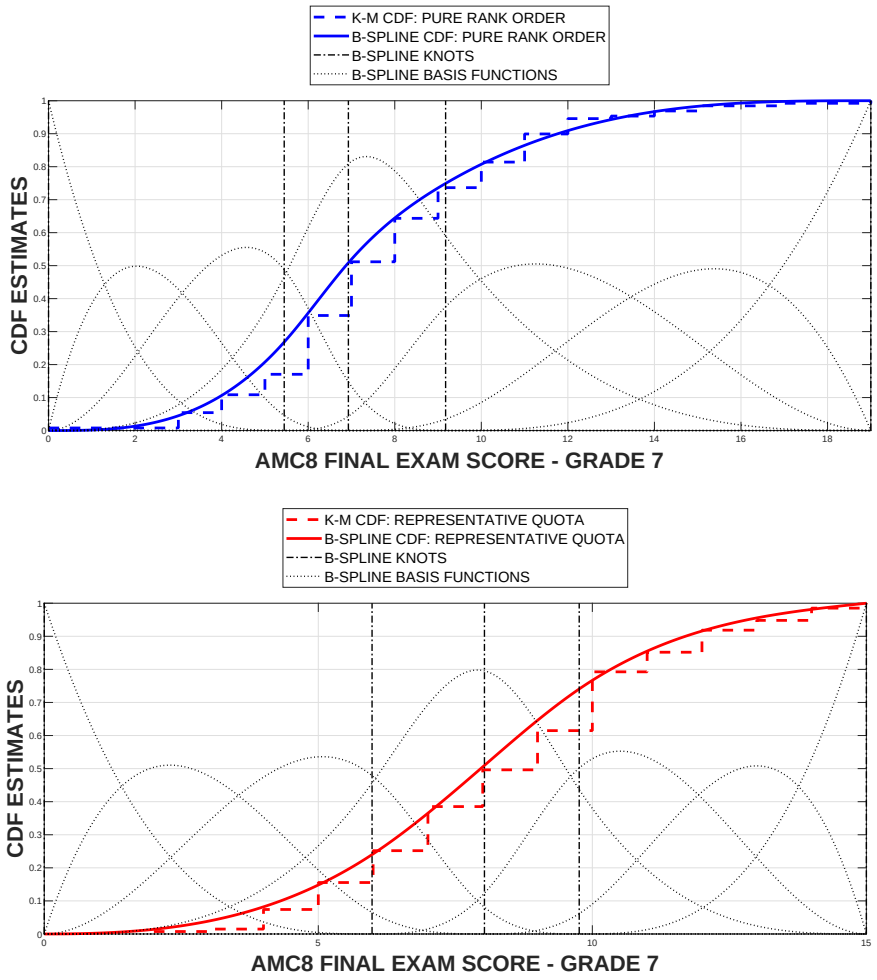
$$\Pr\left[K \text{ optimal}|AIC_{j,K}\right] \approx \exp\left(\frac{\min_{l\in\{K_1,K_2,\ldots,\overline{K}\}} AIC_{j,l} - AIC_{j,K}}{2}\right),$$

which approximates the probability that a given value of K minimizes MISE within a given set of candidates $\{K_1, K_2, \ldots, \overline{K}\}$. For 7^{th} graders we ran our estimator on a set of models with K ranging from a value of 2 (i.e., 5 total basis functions/parameters) to a value of 9 (i.e., 12 total basis functions/parameters), and computed the BIC, AIC, and optimality probability for each candidate model. The results of this model selection process are displayed in Table B.1, which favors $K = 4$ subintervals as statistically optimal in the sense of bias-variance trade-off. We adopt this as the specification for all B-Spline CDFs to be estimated.

Figure B.1 displays the knot locations and B-Spline basis functions for the optimal $K = 4$ model (superimposed on the CDF estimates) for 7^{th}-graders in the *PRO* competition (upper panel) and the *RQ* competition (lower panel). Note that the smoothed B-Spline CDFs are linear combinations of the plotted basis functions, with estimated parameters being the weights on each basis function. The basis functions in turn have shapes that are uniquely determined by the knot vectors through the Cox-de Boor recursion formula. From the picture we see intuitively why B-Splines are so flexible, and why at the same time it is easy to impose shape restrictions on them. Note that the indexing of the basis functions from 1 to 7 follows the left-to-right ordering of their modes, which is why imposing monotonicity of the B-spline CDF is equivalent to ordering the parameter values monotonically as in the GMM constraints. Finally, notice that at the upper and lower bounds of the support there is exactly one basis function that attains a nonzero value (of 1 in both cases) at those points; this is why imposing terminal conditions is a simple matter of imposing equality constraints on the first and last parameter values.

B.2. **PARAMETRIC BOOTSTRAP TEST FOR ORDERING OF TOP EXAM SCORES.** In Section 4.2.4 in the body of the paper we reported the results of a parametric bootstrap test for the ordering of the top scores among 7^{th}-graders across the *PRO* and *RQ* treatments. Predictions (**I**)

FIGURE B.1. 7^{th}-GRADE CDFs, KNOTS, & BASIS FUNCTIONS



and (**II**) state that the best and brightest 7^{th}-graders should perform best under a *PRO* rule when competitive pressure for the best prizes is relatively high. This pattern is borne out in Figures 1 and 2, but as an additional test one could bootstrap the sample maxima from the $(7, PRO)$ group and the $(7, RQ)$ group to see whether the raw difference in the exam score upper bounds $\overline{h}_7^{PRO} - \overline{h}_7^{RQ}$ is statistically above zero. In bootstrapping the raw data, it turns out that the granularity of our raw, multiple-choice exam scores becomes a hindrance for running this test: there is roughly a 10% chance that the bootstrapped sample maxima are *exactly the same*, making it unclear whether to count such observations toward being consistent with the null hypothesis or the alternative hypothesis, or whether to drop them altogether.[27]

[27] If we view the multiple choice exam as a rounded-off version of true underlying math proficiency h, which lives on a continuum, then exact ties in h (without round-off error) would be a zero-probability event.

Therefore, we execute a parametric bootstrap test instead. Whereas a non-parametric bootstrap test simulates observations directly from the empirical CDF (a discontinuous step function), a parametric bootstrap simulates data from a smoothed, parametric representation of the CDF of the data instead. This we do, using the optimized B-Spline CDFs discussed in the previous subsection. More formally, our test procedure is as follows:

(1) For each $s = 1, \ldots, S$ iterations, $S = 10^7$, simulate two samples:

   (a) $\boldsymbol{H}_{7,s}^{PRO} = \left\{ h_{7,s,n}^{PRO} \right\}_{n=1}^{N_7^{PRO}}$, being N_7^{PRO} independent draws from $G_7^{PRO}\left(h; \widehat{\boldsymbol{\alpha}}_7^{PRO}\right)$, and

   (b) $\boldsymbol{H}_{7,s}^{RQ} = \left\{ h_{7,s,n}^{RQ} \right\}_{n=1}^{N_7^{RQ}}$, being N_7^{RQ} independent draws from $G_7^{RQ}\left(h; \widehat{\boldsymbol{\alpha}}_7^{RQ}\right)$.

(2) For each s, compute $\overline{\Delta}_s = \max\left\{\boldsymbol{H}_{7,s}^{PRO}\right\} - \max\left\{\boldsymbol{H}_{7,s}^{RQ}\right\}$

(3) Compute $P = \frac{\sum_{s=1}^{S} \mathbf{1}\left(\overline{\Delta}_s \leq 0\right)}{S}$ (where $\mathbf{1}(\cdot)$ is an indicator function) as the p-value of the null hypothesis that the top student performs weakly best under a RQ rule, against the one-sided alternative hypothesis that the top student performs best under a PRO rule instead (i.e. predictions $(\mathbf{I})$ and $(\mathbf{II})$ are satisfied).

Running this parametric bootstrap test using the optimized B-Spline CDF with $(K+1) = 5$ knots (see previous subsection) results in a p-value of 0.0619, thus rejecting the null hypothesis in favor of predictions $(\mathbf{I})$ and $(\mathbf{II})$.[28]

As a robustness check, we can also execute the above testing proceedure using a kernel-smoothed CDF estimator instead of a B-Spline CDF estimate. For our purposes here, reliable tail behavior is crucial, and standard kernel-smoothed estimators are known to exhibit excessive bias near the bounds of the support. For this reason, we use a boundary-corrected kernel-smoothed CDF estimator developed by Karunamuni and Zhang (2008), which is uniformly consistent on the closure of the support and does not display the tail-bias properties of standard kernel-smoothed estimators.[29] After executing the parametric bootstrap test procedure, but in step (1) simulating samples from boundary-corrected kernel CDFs based on an Epanechnikov kernel funcion and Silverman's automatic bandwidth selection rule instead of using B-Splines, we find a slight strengthening of the test conclusion: the resulting p-value is now 0.0332. Together, these results provide additional evidence that the top 7^{th}-grade student performs better under a PRO rule, relative a RQ rule, as predicted by Cotton, Hickman and Price (2020).

B.3. **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

[28] Re-running this test with any other specification of K between values of 2 and 9 produces some variation in p-value, but under all cases the null hypothesis is rejected at the 10% level.

[29] For an in-depth discussion on boundary-corrected kernel smoothing and its application to applied microeconomics, including a Monte Carlo study comparing its performance with standard kernel-smoothed estimators, see Hickman and Hubbard (2015).

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 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 a different data generating process, say time elapsed outside of learning activity due to work stoppages. 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 $t_j \leq t_{j+1}$ for all $j = 1, \ldots, J$.

(2) Using the sample $\{t_j\}_{j=1}^{J}$, compute an appropriately chosen bandwidth b_1 for a KDE based on a kernel function with support on $[-1, 1]$.[30] Then find the smallest $j_1^* < J$ such that $t_{j_1^*+1} - t_{j_1^*} > 2b_1$. If no such j_1^* exists, then stop; no truncation is needed.

(3) Define initial truncation point $\tau_1 \equiv t_{j_1^*} + b_1$, and compute bandwidth b_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

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

then update the truncation point by $\tau_k \equiv t_{j_k^*} + b_k$, and re-compute bandwidth b_{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+1}^* does not exist (meaning that for the censored sample $\{t_j\}_{j=1}^{j_k^*}$ a KDE is strictly positive everywhere).

[30] 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*.

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 $b_1 = 2.345\widehat{\sigma}_1 J^{-1/5}$ in the first iteration, and $b_k = 2.345\widehat{\sigma}_k j_{k-1}^{*-1/5}$ in the k^{th} iteration ($k \geq 2$), where $\widehat{\sigma}_k$ 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, though choice of the specific kernel is needed to pin down the leading constant on the bandwidth selection rule.

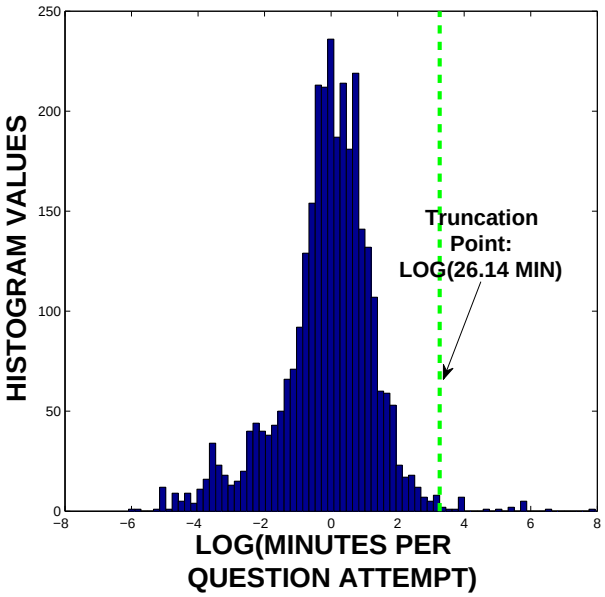
Executing this process on our data leads to a final truncation point of $\tau_2 = 27.81$ minutes per question (the 99.35^{th} percentile of the un-censored sample), after 2 iterations. Figure B.2 displays a histogram of time spent per question, including observations above and below the truncation point. 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.

B.4. **ADDITIONAL FIGURES.** Here we present some additional figures depicting the empirical distributions of investment activities by group and treatment status. In interpreting these figures, one caveat should be kept in mind. Predictions (**I**) and (**II**) only directly apply to the plots in Figures 1 – A.2, since these depict CDFs of exam scores, the variable being directly incentivized within the experimental study. The theory 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 in Figures B.3 – B.4 here.

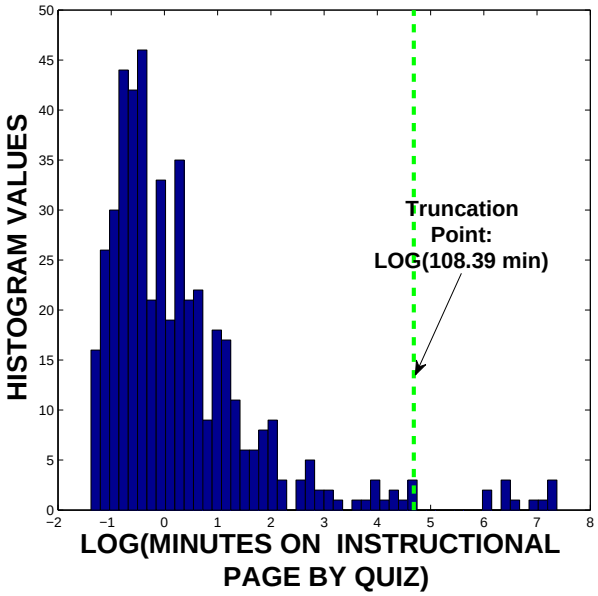
B.5. **CALIBRATING REAL-WORLD, PRE-COLLEGE ACADEMIC INCENTIVES.** We provide some rough calculations here to get a basic sense of the magnitudes of time use incentives for an average high school student studying for college, and how the expected hourly wage within our experiment might compare to it. To do so we calibrate payoffs using market aggregate figures taken from various data sources.

We use data from the Current Population Survey (CPS) from 2003 to calibrate the average net present value (NPV) of lifetime, post-college-age, personal income for workforce participants with different levels of education. We separate income by age for individuals 23–65 and take averages within each age-education group, to get $I_{a,e}$, $a \in \{23, \ldots, 65\}$, $e \in \{cg, sc, hsg\}$, where cg denotes a college graduate with no additional advanced degrees, sc denotes a high-school graduate who enrolled in college but did not graduate, and hsg denotes a high-school graduate with no college experience. For a given value of the annual time discount factor, δ, we compute $NPV_e^{18} = \sum_{a=23}^{65} \delta^{(a-23+5)} I_{a,e}$, where discounting begins 5 years into the future so as to capture the perspective of a recent high-school graduate, at 18 years of age, who is considering going to

FIGURE B.2. TIME TRUNCATION RULE

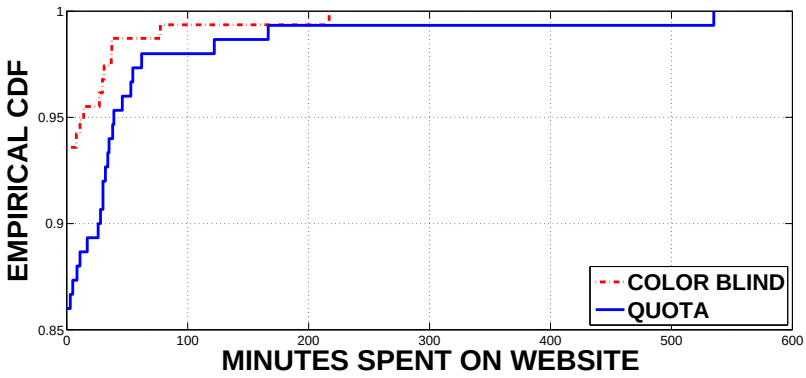


(A) This panel displays a histogram of observed time spent on each question. Each datum in the histogram is a student-question-attempt observation.

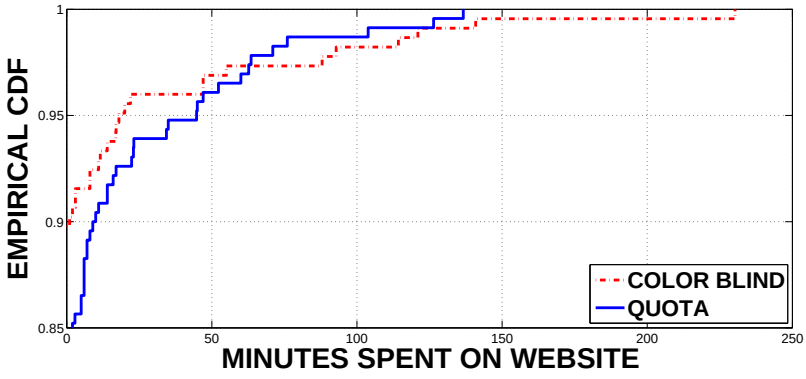


(B) This panel displays a histogram of time per instructional page view. Each datum in the histogram is a student-quiz-attempt observation.

FIGURE B.3. TIME SPENT:
PRO vs. *RQ*



(A) Seventh Graders



(B) Eighth Graders

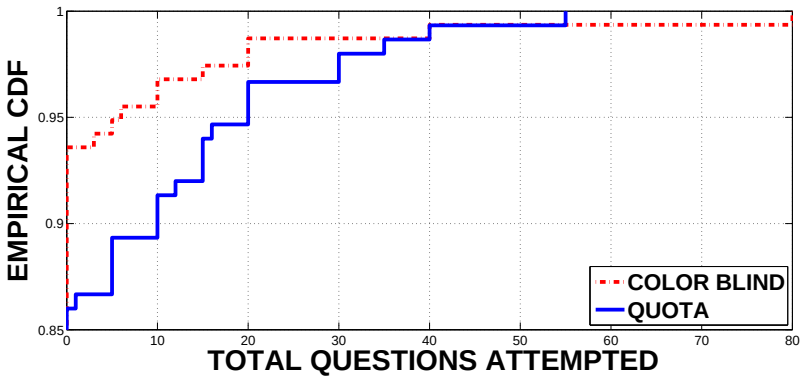
college.[31] We adopt these numbers as a rough estimate of the average gross payoff to the three education options.

As for costs, the College Board estimated that the average cost of attendance during the 2002-2003 academic year was $28,090 at four-year, private, non-profit universities and $12,376 at four-year publics, which we denote by *AnnualCOA*. We shall assume that these same costs prevailed over all four previous years of study.[32] The National Center for Education Statistics reports the 6-year college graduation rate, which we denote by *GradRate*, for the entering class of 1996 (*i.e.*, all those who had graduated by AY2002-2003) as 0.631 for private universities, and 0.517 for

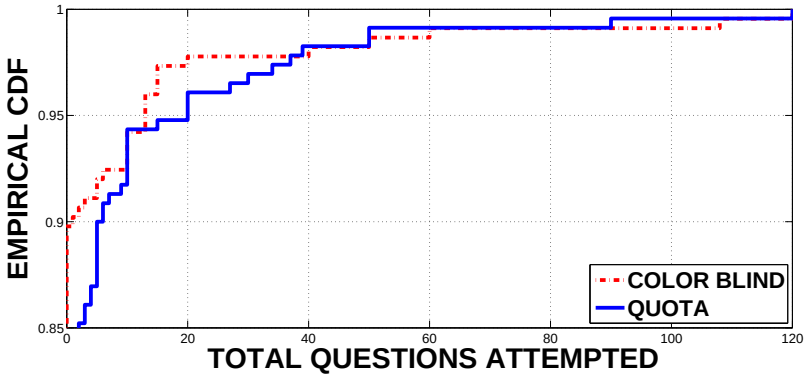
[31]For this rough calculation we omit the comparison during college years as it is unclear what the appropriate comparison would be. College attendees forgo 4 years of labor force income, but most also remain their parents' dependents until graduation, meaning their consumption during college to some degree is still a function of their parents' income. It is also unclear the extent to which recent high school graduates who do not attend college still derive consumption from their parents' income stream. We therefore proceed on the assumption that by age 23 individuals of all education levels become independent adults.

[32]This is a conservative assumption, given that education costs were steadily on the rise during this period. The cost of attendance number includes tuition, fees, room and board, books/supplies, transportation, and other living expenses. Figures taken from Table 1 of report *Trends in College Pricing*, downloaded at http://www.collegeboard.com/prod_downloads/press/cost03/cb_trends_pricing_2003.pdf on April 29, 2016.

FIGURE B.4. QUESTION ATTEMPTS:
PRO vs. *RQ*



(A) Seventh Graders



(B) Eighth Graders

publics.[33] We shall assume that college dropouts incur 2 years of college attendance cost before dropping out.

In order to approximate how many hours, on average, high school students devote to academics, we turn to the Schools and Staffing Survey (SASS), published by the National Center for Education Statistics. SASS reports that in 2003 the average number of hours in the school day for American K-12 schools was 6.6, and that the average number of school days in their academic year was 178.7. If we assume that each school day is coupled with 3.5 hours of homework for high schoolers, then this gives us a measure of $HSHoursWorked = 4 \times 178.7 \times (6.6 + 3.5) = 7,219.48$.[34] If we double the number of daily homework hours to 7 then this number rises by a third. Finally, we put these various measures together for a rough measure of how 18-year-old college enrollees

[33]Numbers taken from *The Digest of Education Statistics*, downloaded from https://nces.ed.gov/programs/digest/d13/tables/dt13_326.10.asp on April 29, 2016.

[34]Figures downloaded from SASS reports at https://nces.ed.gov/surveys/sass/tables/sass0708_045_s1n.asp and https://nces.ed.gov/surveys/sass/tables/sass0708_046_d1n.asp on April 29, 2016.

TABLE B.2. Projections of Effective Hourly Wage for College-Bound High Schoolers

		PUBLIC COLLEGE		PRIVATE COLLEGE	
		HS Homework Hours		HS Homework Hours	
		3.5/school day	7/school day	3.5/school day	7/school day
ANN. DISCOUNT FACTOR	$\beta = 0.92$	\$16.99	\$12.62	\$13.77	\$10.23
	$\beta = 0.96$	\$34.90	\$25.92	\$33.72	\$25.04
	$\beta = 0.98$	\$54.10	\$40.18	\$55.10	\$40.92

are compensated for the time they invested into learning during 4 years of high school:

$$HSHourlyWage = \Big[GradRate \times \left(NPV_{cg}^{18} - \textstyle\sum_{t=1}^{4} \delta^t AnnualCOA \right) + (1 - GradRate) \times \left(NPV_{sc}^{18} - \textstyle\sum_{t=1}^{2} \delta^t AnnualCOA \right) - NPV_{hs}^{18} \Big] \Big/ HSHoursWorked. \tag{B.2}$$

The numerator represents a lump-sum transfer an average college-bound high school graduate would require in order to forgo college, and the denominator normalizes this transfer by the number of total hours spent preparing for college. Since this number will vary by assumptions of discount factor, college option (*i.e.*, public vs. private), and daily homework time inputs, Table B.2 presents various hourly wage calculations covering several different scenarios. A commonly assumed annual discount factor in the macroeconomics literature is $\delta = 0.96$, and a student with these time preferences who studies 3.5 hours per school day will garner an effective return of \$33.72 per hour if she attends a private university, and \$34.90 per hour if she attends a public. Depending on the various assumptions, this number could range from \$10 to \$55 per hour.

At the end of the day, the rough calculations in Table B.2 cannot speak directly to the question of how much a RQ alters expected economic outcomes from competitive human capital investment, and how these changes interact with individual ability and forward-looking behavior. How these various forces balance out in the long run is a challenging question that deserves additional attention from researchers. Answering this question directly using experimental methods would would require an ability to track time use decisions and exogenously shifting market incentives over a period of years or even decades. However, the numbers above suggest that our experimental incentives may not be entirely un-representative of the magnitudes of incentives that high-school students face on a regular basis when making decisions about leisure-study tradeoffs with an eye toward a college education.

APPENDIX C. STUDENT INFORMATION SHEETS

Figures B.5–B.10 below display examples of the information sheets given to test subjects explaining the distribution of their competitors and the distribution of the prizes at stake.

FIGURE B.5. PRO Competition Student Information Sheet Example (front)

BYU AMC 8 Math Contest

On Tuesday November 13th you will have the chance to the take the AMC 8 during school. The AMC 8 is a nationally recognized math exam that takes 40 minutes and involves 25 multiple choice questions. We are providing some incentives for students to prepare for and do well on the exam.

You have been randomly assigned to the green group. You will be competing with 145 other 7th and 8th graders at various middle schools in Utah County. They will all be taking the same test on the same day as you. Among these 145 students we will be giving out the following cash prizes:

Prizes will be awarded to the top 48 students in the following way:

The top 3 scores will receive $34,

the next 3 scores will receive $32,

the next 3 will receive $30, and so on, .

A complete table of prizes is provided on the back of this sheet.

We will score the tests the day after the exam and deliver the prizes to your school within the next week. The highest possible score on the AMC 8 is 25. The following table provides information for the 7th and 8th graders who took the practice test. The average score was 7.0 for 7th graders and 8.7 for 8th graders.

Performance of 7th and 8th graders (based on the practice exam)
10% of students scored 12 or higher
20% of students scored 11 or higher
30% of students scored 10 or higher
40% of students scored 9 or higher
50% of students scored 8 or higher
60% of students scored 7 or higher
70% of students scored 6 or higher
80% of students scored 6 or higher
90% of students scored 4 or higher

We have set up a website with practice problems that we gathered from past years exams of the AMC 8. Practicing these problems will improve your performance on the AMC 8 and increase your chances of winning one of the larger prizes. KEEP IN MIND THAT OTHER STUDENTS COMPETING FOR PRIZES WILL BE PRACTICING TOO, SO DON'T GET LEFT BEHIND! You can access that website at: **http://byuresearch.org/math/**

FIGURE B.6. PRO Competition Student Information Sheet Example (back)

COMPLETE PRIZE TABLE

1st-3rd place = $34,

4th-6th place = $32,

7th-9th place = $30,

10th-12nd place = $28,

13rd-15th place = $26,

16th-18th place = $24,

19th-21st place = $22,

22nd-24th place = $20,

25th-27th place = $18,

28th-30th place = $16,

31st-33rd place = $14,

34th-36th place = $12,

37th-39th place = $10,

40th-42nd place = $8,

43rd-45th place = $6,

46th-48th place = $4.

While these prizes may seem large, keep in mind that the reward to doing well on some math tests such as the ACT or SAT can be thousands of dollars in terms of scholarships or future earnings.

FIGURE B.7. 7th Grade RQ Competition Student Information Sheet Example (front)

BYU AMC 8 Math Contest

On Tuesday November 13th you will have the chance to the take the AMC 8 during school. The AMC 8 is a nationally recognized math exam that takes 40 minutes and involves 25 multiple choice questions. We are providing some incentives for students to prepare for and do well on the exam.

You have been randomly assigned to the yellow group. You will be competing with 40 other 7th graders at various middle schools in Utah County. They will all be taking the same test on the same day as you. Among these 40 students we will be giving out the following cash prizes:

Prizes will be awarded to the top 14 students in the following way:

The top score will receive $34,

the next score will receive $32,

the next will receive $30, and so on,

A complete table of prizes is provided on the back of this sheet.

We will score the tests the day after the exam and deliver the prizes to your school within the next week. The highest possible score on the AMC 8 is 25. The following table provides information about scores of the 7th graders who took the practice test.

Performance of 7th graders (based on the practice exam)
10% of students scored 10 or higher
20% of students scored 9 or higher
30% of students scored 8 or higher
40% of students scored 7 or higher
50% of students scored 7 or higher
60% of students scored 6 or higher
70% of students scored 6 or higher
80% of students scored 5 or higher
90% of students scored 4 or higher

We have set up a website with practice problems that we gathered from past years exams of the AMC 8. Practicing these problems will improve your performance on the AMC 8 and increase your chances of winning one of the larger prizes. KEEP IN MIND THAT OTHER STUDENTS COMPETING FOR PRIZES WILL BE PRACTICING TOO, SO DON'T GET LEFT BEHIND! You can access that website at: **http://byuresearch.org/math/**

FIGURE B.8. 7th Grade RQ Competition Student Information Sheet Example (back)

COMPLETE PRIZE TABLE

1st place = $34,

2nd place = $32,

3rd place = $30,

4th place = $28,

5th place = $26,

6th place = $24,

7th place = $22,

8th place = $20,

9th place = $18,

10th place = $16,

11st place = $14,

12nd place = $12,

13rd place = $10,

14th place = $8.

While these prizes may seem large, keep in mind that the reward to doing well on some math tests such as the ACT or SAT can be thousands of dollars in terms of scholarships or future earnings.

FIGURE B.9. 8th Grade RQ Competition Student Information Sheet Example (front)

BYU AMC 8 Math Contest

On Tuesday November 13th you will have the chance to the take the AMC 8 during school. The AMC 8 is a nationally recognized math exam that takes 40 minutes and involves 25 multiple choice questions. We are providing some incentives for students to prepare for and do well on the exam.

You have been randomly assigned to the blue group. You will be competing with 100 other 8th graders at various middle schools in Utah County. They will all be taking the same test on the same day as you. Among these 100 students we will be giving out the following cash prizes:

Prizes will be awarded to the top 32 students in the following way:

The top 2 scores will receive $34,

the second 2 scores will receive $32,

the next 2 will receive $30, and so on,

A complete table of prizes is provided on the back of this sheet.

We will score the tests the day after the exam and deliver the prizes to your school within the next week. The highest possible score on the AMC 8 is 25. The following table provides information among the 8th graders who took the practice test.

Performance of 8th graders (based on the practice exam)
10% of students scored 12 or higher
20% of students scored 11 or higher
30% of students scored 10 or higher
40% of students scored 9 or higher
50% of students scored 9 or higher
60% of students scored 8 or higher
70% of students scored 7 or higher
80% of students scored 6 or higher
90% of students scored 5 or higher

We have set up a website with practice problems that we gathered from past years exams of the AMC 8. Practicing these problems will improve your performance on the AMC 8 and increase your chances of winning one of the larger prizes. KEEP IN MIND THAT OTHER STUDENTS COMPETING FOR PRIZES WILL BE PRACTICING TOO, SO DON'T GET LEFT BEHIND! You can access that website at: **http://byuresearch.org/math/**

FIGURE B.10. 8th Grade RQ Competition Student Information Sheet Example (back)

COMPLETE PRIZE TABLE

1st-2nd place = $34,

3rd-4th place = $32,

5th-6th place = $30,

7th-8th place = $28,

9th-10th place = $26,

11st-12nd place = $24,

13rd-14th place = $22,

15th-16th place = $20,

17th-18th place = $18,

19th-20th place = $16,

21st-22nd place = $14,

23rd-24th place = $12,

25th-26th place = $10,

27th-28th place = $8,

29th-30th place = $6,

31st-32nd place = $4.

While these prizes may seem large, keep in mind that the reward to doing well on some math tests such as the ACT or SAT can be thousands of dollars in terms of scholarships or future earnings.