Identifying and Nurturing Math Talent
Evidence from Tamil Nadu*

Phi Adajar, Esther Duflo, Glenn Ellison, Sara Fisher Ellison, and Harini Kannan†

March 2026

PRELIMINARY: NOT FOR DISTRIBUTION

This paper presents evidence from a randomized controlled trial designed to test how to best identify and support gifted students from disadvantaged backgrounds. To do so, we offered access to an advanced online prealgebra course and an internet-enabled tablet to 7th-grade public-school students from Tamil Nadu, whom we identified as having high potential using several alternative strategies. Our primary finding is that we were able to successfully identify students who succeeded in challenging coursework originally designed as enrichment material for students with much stronger math preparation. We randomly assigned identification methods, which allowed us to investigate their relative effectiveness. We find that a combination of administrative scores and engagement with a booklet containing an extract of the textbook best predict engagement and success with the course material. We additionally cross-randomized treatments to evaluate program design (including group and individual tutoring, incentives for students, and access to computer labs), and find that the simplest-to-scale version without tutoring led to the highest engagement.

*We are extraordinarily grateful to the partners that made this research possible: the Art of Problem Solving, for access to course materials and for tremendous technical and logistical assistance, and in particular, Richard Rusczyk, Yating Liu, and Jo Welsh; and the Tamil Nadu School Education Department, for their collaboration in implementing this project and in sharing administrative data. Special thanks to Sadhana Babuji, Kannan Narayanaswamy, and Akhilan Rengaswamy for outstanding research assistance; Veena Jayaram, Nitin Kanakaraj, Murali Mallikarjunan, and Mustufa Patel for research management; our 57 tutors, for their dedication to supporting students; and the over 100 members of our field team, who collectively visited over one thousand schools over the course of this project. Funding was provided by the Agency Fund, Founders Pledge, and the J-PAL Science for Progress initiative. This project was approved by MIT's Committee On the Use of Humans as Experimental Subjects (Protocol Nos. 2308001072, 2408001374, and 2410001443) and the IFMR Institutional Review Board (Protocol Nos. 10705, 10708, 10713, 10740, 10775, 10776, 10795, 10822, 10854, 10916, 10987, 11013, and 11034). This experiment was pre-registered as RCT #14090 in the AEA RCT Registry.

†padajar@mit.edu, eduflo@mit.edu, gellison@mit.edu, sellison@mit.edu, and harini.kannan@ifmr.ac.in.

1 Introduction

India produces many successful engineers and scientists—according to some estimates, 23% of tech workers in Silicon Valley who have an advanced degrees were born in India—but most come from families who had the means to send their children to private schools and tutoring centers. By some estimates, less than 5% of students who earn qualifying scores on the Joint Entrance Exam for the elite Indian Institutes of Technology (IIT) attended a government (public) school, even though 60% of Indian students attend such schools. This sparsity of less-advantaged Indian students at the top of the distribution of achievement is reflected in international comparisons: when 15-year-old students in Tamil Nadu government schools participated in the PISA in 2009, the 95th-percentile student achieved the same score as the median student in the United States.[1]

This problem is not unique to India: top students in math and science disproportionately come from rich countries. For example, Hoyos (2019) finds a large regional gap at the International Mathematics Olympiad, with only 0.1% of gold medalists being from Africa.[2] There is no reason to think the underlying distribution of raw talent is different, so the scarcity of top students must reflect identification and training. A large literature has highlighted the rigidity of the school curriculum in many developing countries, including India. These curricula are often too advanced for the majority of the students in the class, leaving many children behind. It also seems plausible that a rigid curriculum will not be adapted to some very talented students, who would at best not flourish, and at worst get bored and disengage. This is particularly likely in an environment where most middle- and upper-class students, especially in urban areas, attend private schools or complement their formal education with tutoring outside of school. Less-advantaged students who only attend government schools are simply not expected to perform at a high level, and there is very little structure to help them in doing so.

A failure to develop high-ability students in government schools will lead to a significant misallocation of talents, where many potential engineers, scientists, and entrepreneurs never get to realize their potential. Nevertheless, while there is an active literature on talent programs in the US, the research on education in developing countries has mostly focused on how to change the curriculum to better serve struggling students.

In this paper, we take a first step to fill this gap. Specifically, we investigate whether the availability of advanced online material makes it possible to nurture talented math students in resource-poor settings: Can they be identified? Can they be given access to the infrastructure they need to reliably access the material? Do they have the time, stamina, and ability to follow the material even though it was designed for motivated, well-supported children in wealthy countries? Can a program be designed such that it can be rolled out at sufficient scale to identify a fraction of very talented

[1] See Walker (2011) for the specific results; India has not agreed to participate in the PISA since, arguing that the test does not reflect the skills taught in the Indian school system.

[2] See also Steegh et al. (2019).

children, and allow them to succeed with the material?

To answer these questions, we set up a randomized controlled trial in partnership with the government of Tamil Nadu and a provider of high quality online math content, *The Art of Problem Solving*, which donated the content and student access for the trial. The experiment covered 2,490 promising 7th-grade students from disadvantaged backgrounds attending government schools in Tamil Nadu. The main treatment, provided to 1,605 of them (randomly assigned), was a self-paced online enrichment course in prealgebra, along with a tablet and an internet connection. In addition, we introduced two additional sources of randomization. First, we randomly varied the criteria used for selecting participants. We therefore can study different methods of identifying promising students and comment on both scalability and effectiveness. Second, within the treatment group who all received the enrichment course, we varied the details of the implementation to study whether additional support is critical to student engagement and success with the program.

Tamil Nadu is one of India's wealthiest states, and is proud of a strong math, science, and engineering culture. At the same time, the state recognizes the challenges that poorer students face to attain academic excellence, and has started several initiatives in response. First, students are regularly tested in school, allowing for ongoing monitoring. Second, the state runs several schemes to support poor talented children, including specialized high schools and scholarships for poor students who have done well in grade 10. However, state officials recognize that there remains considerable scope for improvement and were interested in experimenting with a program that provides access to online high quality education to talented children.

Despite the enthusiasm of our government partners and schools, and the availability of a very valuable pedagogical resource, the setting presented a number of challenges which made it less than evident that the program would succeed. The course material, designed for advanced 6th or 7th graders in the US, may be too challenging for students (and especially those with limited English proficiency). Most parents have at most a 10th grade education, so it is not clear if they or anyone in the community will be able to provide support. Most of the students lack regular internet access. This made it important to provide tablets to allow for access, but also meant that those tablets could open access to distracting materials. Additionally, students have a lot of competing demands on their time, and may not have the time or motivation to devote to the program without any direct reward if they do.

Given this context, our most important finding is perhaps that, despite these substantial challenges, roughly 40% of the students *invited* to participate in the AOPS program did engagee with most of the 366 components of course content—they clicked on the links that show the readings, they started the videos, they displayed the problems, and so forth. We define a stringent multi-factor performance threshold we refer to as "successful course completion", which includes passing at least 90% of the course's embedded quizzes. Twenty percent of students invited to participate

reached this level in a few months, a high level for a self-paced course.

We present evidence of the relative success of students selected by different methods over which we randomized. Methods included administrative (standardized) test scores, teacher recommendations, engagement with a printed version of Chapter 1 of the course textbook, performance on a quiz we administered, and a measure of their "grit." Our most important finding here is that student engagement with the printed excerpt is a powerful predictor of later online course engagement and performance.

Next, we present evidence on program design features, which were fully cross-randomized with the selection methods. Some students were offered tutoring (individual or group). Some students were offered access to a high-tech lab in the schools where they could work together. Some students were provided incentives: they were explicitly told that their tablets would be removed if they did not progress through the course (and to differentiate the incentive effect from the direct impact of removing the tablets from non performing students, we did in fact take away the tablets for a randomly selected subset of students).

The most important result here is that tutoring support is not a necessary feature of the program. This is very encouraging for the possibility of scaling up the program. However, a surprising result is that students assigned to either group tutoring or individual tutoring made significantly *slower* progress than those who were not provided any tutoring support. We present some explanations to help understand this.

We pre-registered some heterogeneity analysis in program effectiveness. Among other dimensions, we look at differences by gender, and note that two-thirds of our participating students, selected with a gender-blind procedure, are female, and that the female students show greater course engagement and learning. This finding is striking given the large gender gap—favoring men—within STEM-related higher education and professions in India, and suggests this may reflect differential treatment of girls and young women: promising boys may be sent to private school, for example, and thus not appear in our sample, while promising girls stay in government schools.

Finally, we present evidence on the impact of the program on broader learning. We know from the embedded quizzes that many of our treatment students mastered challenging material, but are also interested in whether students improved their general problem-solving skills. To assess this, we administered a very challenging endline exam to both treatment and control students. Our exam was, unfortunately, overly challenging for our study population—the average score is just 11%—which limits our ability to detect moderate gains. Nonetheless we find significant positive effects for the treatment group in specifications that drop students who cheated. The measured gains in the ability to solve problems at the level we tested are, however, modest to date. Larger differences could emerge when treatment students take additional courses, or if the test covers a broader array of skills.

Our final set of results addresses potential unintended consequences of our intervention. One might be concerned that providing internet access and devices might crowd out learning by enabling video streaming, game playing, and so forth. Students returned the tablets to us at the end of the program, which made it feasible to gather unusually detailed data on tablet use with course and school outcomes. We observe large amounts of time spent streaming video for entertainment, but we do not find evidence of net negative effects on academic performance for students who spent more time video streaming.[3] We provide tentative evidence that net effects are positive for students whose tablet use indicates engagement with the enrichment content.

Our paper touches on the large and active literature on online education, education in developing countries, and gifted education. We will only sketch their broad contours, referring to survey articles when appropriate, and indicate where this paper fits in. Escueta et al. (2020) is an excellent pre-COVID survey of the effects of educational technology, including online content.

Our focus is different from much of the EdTech literature, in that we focus on high-potential students. As such, it is also part of the literature on gifted education. A number of papers have used regression-discontinuity designs and other methods to estimate causal effects of elite schools and other programs serving gifted students, e.g. Abdulkadiroğlu et al. (2014), Dobbie and Fryer (2014), Bui et al. (2014), Card and Giuliano (2024), Cohodes (2020), and Ellison and Pathak (2025). This paper studies a more light-touch intervention that does not replace traditional education but supplements it for talented students. Our focus on the best way to *identify* these talented students also appears to be novel.

Our paper sits in the literature on missing talent and achievement gaps. Others have found evidence of missing talent in other educational settings. For example, Hoxby and Avery (2012) report that the vast majority of very high-achieving students who are low-income do not apply to any selective college, and Ellison and Swanson (2016) report that schools serving similar (and relatively well off) students in the US produce high-achieving math students at very different rates, suggesting missed talent. Our finding that a significant population of students in an under-resourced setting successfully engage with material much more challenging than they are exposed to in school is both another example of missing talent and a demonstration that particular students who would have been missed can succeed when given an opportunity. One can also interpret literature on the STEM gender gap as indicative of missed talent.[4] Our findings on the gender mix are an important piece towards better understanding the mechanics of the STEM gender gap in India. Our finding that so many of the top math students in government schools are girls is an interesting contrast with what has been reported in so many other settings.

[3] We heard anecdotally from a number of students that other members of their families were using the tablets to access the internet. As a result, we suspect that the time-use data over-states student accessing of other content.

[4] See, for instance, Cimpian and King (2024); Ellison and Swanson (2010).

As we mentioned, the literature on identifying and nurturing talented students in developing countries is sparse, but earlier findings from other educational interventions, especially those in developing countries, helped to inform our design.[5] Previous studies have found evidence that self-paced classes and tutoring help students across the ability distribution.[6] An interesting example is Carlana and Ferrara (2021), which used an RCT to examine a Covid-era tutoring intervention targeted at vulnerable students who were in danger of falling behind during remote learning. They found significant, positive effects of tutoring for these students. We complement this literature in examining tutoring effects for high-ability, high-motivation students, and our findings that tutoring seems to have held back student progress raise questions about what aspects of our environment contribute to our contrasting result. We also contribute to a small literature on talent identification.[7] To complement prior work, we explicitly randomize over many potential inclusion criteria to shed light on the most effective ones, and use engagement with a mini-version of the course as a potential criterion. As a result, we believe that our study makes an important and unique contribution on how to best identify students who might benefit from interventions like ours.

Finally, our paper is also related to the literature on the unintended consequences of providing internet-enabled computers to students. Malamud and Pop-Eleches (2011) report that home computers provided to disadvantaged secondary-school students in Romania reduced school grades by one-quarter to one-third of a standard deviation, and time-use data provided a clear potential mechanism: students reported spending more time playing computer games and less on homework and reading. Malamud et al. (2019) examine a 2011 RCT that provided disadvantaged Peruvian 3rd- to 5th-grade students with both home Internet access and instruction in how to use laptops to access educational resources. They find a small increase in time spent on games, social media, and so forth, and no effect on test scores.[8] Some features of our setting that complement the existing literature are our focus on high-achieving students, and that we are examining the effects of mobile devices in an era when social media and short-form videos have made concerns about the effects of online access even more salient. In line with the existing literature, we find that the tablets were extensively used for non-educational activities, but, similar to the latter paper, do not find evidence of a reduction in scores on end-of-year exams in any subject.

[5] Moreira (2019) notes that recognition provided to high-achieving students can also provide spillover benefits to classmates.

[6] See, for instance, Carlana and Ferrara (2021); Muralidharan et al. (2019); Koedinger et al. (1997); Albo et al. (2025).

[7] See Saavedra and Buitrago (2025) for a randomized intervention to identify potential Math Olympians in public schools in Colombia.

[8] Beuermann et al. (2015) also find no test-score effect in an RCT that provided internet-enabled laptops to second-grade students in Peru.

2 The Environment and Experiment

2.1 Student population

We carry out our experiment in the Indian state of Tamil Nadu, which enrolls just over 1,000,000 7th-grade students. About 460,000 of these students attend government schools, which is the population we study. We operated with cooperation from the Tamil Nadu School Education Department, who provided us with school- and student-level data, as well as access to the schools, the headmasters, and the teachers. Government schools serve a relatively disadvantaged population compared to private and government-aided schools.

We first narrowed our sample to ten relatively underprivileged districts with 112,630 7th-grade students and 3,494 government schools.[9] Of students in these districts, 38.1% of them are a member of a "Scheduled Caste" or "Scheduled Tribe" (the most disadvantaged groups), compared to 21% of the population-at-large. Remaining students largely belong to a "Most Backward Class" (38.6%) or "Other Backward Class" (21.1%). Note that all of the these designations, comprising 97.8% of our student population, qualify for affirmative action programs in employment and education.

We then narrowed our population further to schools which reported multiple high-scoring students or at least one very high scorer on administrative (standardized) math exams. That left us with 1,441 schools and 75,533 7th grade students. We then selected 9,709 initial participants from this group that met one of several criteria: sufficiently high 6th-grade administrative score, teacher identification as a "top student," or teacher recommendation as a student with high potential not reflected in administrative scores. More details about this initial process for selecting students for further screening are given in Appendix A.

In our final screening stage, we provided the 9,709 students with a printed booklet containing an excerpt of the companion textbook for the online math course we would make available in the treatment. We asked the students to work through the printed booklet, after which we administered a quiz and survey to each student. The process by which we selected promising students from the 9,709 for participation in the RCT was randomized across schools, and is described fully in Section 2.2.1. In total, we invited 2,490 students across 867 schools to participate in the RCT, representing 3.4% of students within the set of 1,441 schools.

Table 1 presents demographics of our student population in the ten districts, the ones who are selected for the final screening, and our final RCT sample. At each successive stage, our sample becomes relatively less disadvantaged. Recall, however, that our population in the ten districts is disadvantaged at baseline. One interesting note is that about two-thirds of our final sample are

[9] The ten districts are Chennai, Cuddalore, Dharmapuri, Kancheepuram, Nagapattinam, The Nilgiris, Tirunelveli, Tiruvanamalai, Vellore, and Vellupuram. We concentrated our experiment in a small number of districts to reduce travel time to schools for our staff.

girls, despite gender not being a component of our selection.

Table 1: Demographics of RCT sample.

	(1) All Students	(2) Selected for Screening	(3) Selected for RCT
Male	0.484	0.371	0.332
Scheduled Caste/Tribe	0.381	0.279	0.236
English-medium	0.330	0.367	0.443
Father's Employment Type			
Daily Wage	0.781	0.720	0.654
Private	0.081	0.102	0.135
Government	0.009	0.020	0.028
Father's Education Level			
5th Grade and below	0.179	0.126	0.113
8th Grade	0.250	0.219	0.181
10th Grade	0.404	0.414	0.405
12th Grade and above	0.167	0.242	0.301
6th-grade Math Score	66.086	83.223	89.522
6th-grade Math Score (in-school z)	0.000	1.398	1.688
Num. Students	112,630	9,709	2,490
Num. Schools	3,494	1,355	867

Table includes means of various demographic and academic variables for all students in our initial consideration set, all students who were selected for screening, and all students who were included in our RCT sample (passed final screen). Means are reported for the set of students with non-missing values of the relevant variable.

2.2 Treatments

The objective of our broader project is to study the effectiveness of interventions, such as providing challenging online math courses, in nurturing high-level math talent. We believed that starting with math instruction before algebra was most appropriate given the level of math skill that our student population starts with. We chose the self-paced *Prealgebra 1* online course offered by the Art of Problem Solving, as providing both high quality content and an engaging format.[10] It is described further in Section 2.3. Course access was offered for six months, along with a corresponding physical textbook and a tablet with cellular data access. Students were placed in WhatsApp groups to send periodic reminders and announcements and to facilitate connection with other nearby students.

[10]The Art of Problem Solving generously donated course enrollments to our treatment students for this study.

Treatment was randomized at the school level, and each school was treated with probability two-thirds, resulting in 1,605 treated students.[11] We test balance by regressing baseline covariates on treatment status in Table B.1, finding no evidence of imbalance. Treatment began in mid-January 2025 and ran through July 2025. Students and parents likely viewed selection for this program as an honor. In an effort to mitigate the effect of this recognition, control students also received certificates for their high achievement. To limit the extent to which selection for the program might change teachers' behavior towards selected (or non-selected) students, treatment students' teachers received no updates about their progress, nor benefits of any form from our research team or the School Education Department.

Our experiment included cross-randomization with several subtreatments, designed to evaluate different aspects of student selection and program design.

2.2.1 Selection subtreatments

One of the primary questions we wanted to investigate was how to identify students for a long-term intervention who would engage with the material and benefit from the program. To that end, we designed eight selection methods and randomized them at the school level. The selection methods used different combinations of four criteria that we thought could be useful signals for identifying well-matched students: *Admin*, *Booklet*, *Quiz*, and *Grit*.

The *Admin* measure is constructed from an administrative test score: an end-of-year math exam in 6th grade.[12] While highly scalable, the test's low ceiling could make distinguishing above-average from highly-talented students difficult. For our *Booklet* measure, we offered students a physical booklet containing Chapter 1 of AoPS *Prealgebra* and asked their teachers to photograph the work they had done in the booklet after one month. We generated a score for each student based on problem attempts and other work in the booklet as well as checking answers on an online portal. This measure was intended to capture interest in, not mastery of, the material. Our *Quiz* measure is based on a quiz we administered after the students had had the booklet for a month. It was designed by our research team to test students' knowledge of the material from the booklet and problem-solving ability.[13]. Finally, we assessed students' *Grit* using a questionnaire developed

[11] Being part of the "treatment" group and experiencing the program described above are not perfectly correlated. 13 of the 1,605 "treatment" students declined to sign up for AOPS and hence did not actually experience the program. Meanwhile, 22 of the 885 "control" students received tablets and were signed up for the course, due to a combination of mistakes by field workers and our responses to a few headmasters who insisted with sufficient vehemence that a particular student be included. Hence, all estimates of "treatment effects" in the paper should be thought of as intent-to-treat estimates.

[12] We used both within-school z-scores and absolute scores to account for grade inflation. This conversion could be motivated by a model in which the true distribution of scores is identical across schools, but a score-inflation process involves an affine transformation that differs across schools.

[13] The quiz ended up being less useful than we had hoped. While reviewing the scores, we discovered that 70% of the students had likely cheated, copied answers, or had their answers copied. We replaced the scores of these students with their predicted quiz score given their 6th-grade test scores, but only if doing so led to a decrease in their quiz score.

by Duckworth et al. (2007) and a small, unfamiliar puzzle.[14] Table 2 provides statistics on these measures among students selected for the final screening and for those chosen for the RCT. The students selected for the RCT have higher mean values of each criterion.

Table 2: Academic characteristics of RCT sample.

	Selected for Screening		Selected for RCT	
	Mean	SD	Mean	SD
Admin (z)	0.00	1.00	0.52	0.77
Quiz Score (%)	48.0	24.3	62.7	18.1
Booklet (z)	0.00	1.00	0.65	0.99
Grit (z)	0.00	1.00	0.31	1.05
Observations	9,709		2,490	

Table includes means of various academic variables for all students who were selected for screening and for all students who were included in our RCT sample (passed final screen). Means are reported for the set of students with non-missing values of the relevant variable.

We fully specified eight potential methods for selecting students using these criteria and randomized which method was used at the school level. Four of the methods involved selecting students on the bases of a single criterion. Each of these methods was used with 10% probability. The other four methods used combinations of criteria: *Admin+Booklet* (10%) , *Booklet+Quiz* (10%), *Admin+Quiz* (20%), and *Admin+Booklet+Quiz* (20%).[15] We set cutoffs for each selection method before conducting the randomization so that each method would select approximately 2,400 students if used in all schools; pairwise balance tests are presented in Table B.6.

The randomization of schools into selection methods lets us estimate differences in expected course outcomes across selection methods via simple comparisons of sample means. These will be consistent estimates of causal effects under very general assumptions, e.g. even if student outcomes affected by which of their peers are selected to participate. Although the set of students included in the RCT is a selected sample, the selection is purely on observed variables, so under the assumption that effects are solely a function of individual characteristics we can also estimate what course outcomes might have been under counterfactual selection methods by running regressions of course outcomes on the criteria that were used in the selection. Part of the motivation for the specific set of methods used was that we wanted to ensure that there would be substantial het-

[14] We asked the traditional bridge-and-torch problem and constructed the measure based on the amount of time spent, the number of tries, and whether the answer was correct.

[15] The skew toward procedures incorporating the quiz scores reflects the fact that we expected that it would be a good measure of student ability, and that we failed to anticipate the prevalence of cheating.

erogeneity in observables in the RCT sample to improve the power of such a regression. Table 2 shows that substantial heterogeneity is present.[16]

2.2.2 Design subtreatments

Another central question we wanted to investigate was the importance of support features, such as tutoring and computer labs. To investigate how these and other elements of program design affect student engagement, treatment students were cross-randomized at the school level across three sets of subtreatments as described in Table 3. Pairwise balance tests for each subtreatment are presented in Table B.3, Table B.4, Table B.5; we find no evidence of systematic imbalance.

Table 3: Description of design subtreatments.

Subtreatment Set	Subtreatment	Percent of Schools
Tutoring	Individual tutoring	12.5%
	Group tutoring	37.5%
	Control (no tutoring)	50%
Computer labs	Access	25%
	Control (no access)	75%
Incentives	Enforced progress requirement	33%
	Unenforced progress requirement	33%
	Control (no progress requirement)	33%

The first set of subtreatments evaluated the provision of tutoring support from STEM students and graduates of elite universities in Tamil Nadu. While past research found benefits of tutoring in a variety of settings,[17] it was unclear whether we would find similar results here. On the one hand, students in the right tail of talent may have sufficient self-motivation that tutoring is not necessary. On the other, our coursework is designed for high-achievers, but our sample of "advanced" students have the same mathematics knowledge as the median American student. The 37.5% of students assigned to group tutoring attended virtual one-hour sessions with 8–10 students. In addition, about 12.5% of students were assigned to a one-on-one tutor so that we could investigate whether this personal attention could have an impact different from group tutoring. These tutors met for 15 minutes weekly with 10 different students, offering assistance with any topics and providing encouragement. The remaining 50% of students did not have a tutor. Students in all subtreatments could also receive (and provide) help with material by posting (and answering) questions on AoPS' message boards. Students receiving tutoring were placed in WhatsApp groups with their tutor. To provide similar community interaction, students in the non-tutoring subtreatment were also partitioned into WhatsApp groups with about 80 other students from their district.

[16] Table B.2 includes academic characteristics for students selected using each criterion, showing that our various selection methods increased this heterogeneity.

[17] See, for instance, Kraft and Lovison (2024); Angrist et al. (2023); Carlana and Ferrara (2021).

The second set of subtreatments involved an additional way for students to access the course. In recent years, the School Education Department has invested in "Hi-Tech Labs" (HTLs) throughout government schools, providing internet-enabled computer labs for students. Such labs may be useful for helping overcome low engagement with tablets seen in other studies.[18] Schools were randomized into this subtreatment with 25% probability. We worked with school administrators to allow students to work on AoPS in the computer lab after school at least once per week. Most schools randomized into this subtreatment had multiple students, so the students could work together in the lab, facilitating collaboration and peer-to-peer teaching. Some schools had a single student, though, in which case the benefit of this treatment could be a regular work time and a larger screen, potentially mitigating low tablet engagement.

All students received regular messages with progress reports and reminders to engage with the course. The final set of subtreatments was designed to test the impacts of additional messaging and incentives, with the goal of increasing completion rates and reducing tablet misuse. Two-thirds of students were told that if they did not make progress in the course at a sufficient rate, their tablets would be taken away. One third did not. Within the former group, we varied enforcement: half of the students who received the messages would eventually have the tablets taken away if they were not making enough progress, while the other half would not. Again, subtreatments were randomized at the school level. In practice, the messages had not yet progressed from vague to specific by the time schools closed for summer vacation, and no tablets were collected until four months into the program. We eventually collected tablets from approximately 100 students (i.e. about 20% of students in the subtreatment); in theory, these students could earn their tablets back by making progress on the course without the tablet, but none did.

2.3 Mathematics course

Our experiment uses the self-paced *Prealgebra 1* course offered by the Art of Problem Solving (AoPS). AoPS is a private educational firm that began in the early 1990s with two textbooks designed for US middle- and high-school students preparing for math competitions. Since then, the company has developed a nine-volume textbook series, ranging from *Prealgebra* to *Calculus*. These textbooks cover middle-school and high-school material on a subject-by-subject basis, but emphasize problem-solving and include problems that are much more challenging than those found in standard textbook series. AoPS offers synchronous online classes, self-paced online classes, and in-person classes at after-school enrichment programs. These courses serve many students preparing for math competitions and are popular with parents who want to supplement the education their children receive in their (typically elite) schools. AoPS touts that almost all of the students who have represented the US at the International Math Olympiad have taken their courses.

The *Prealgebra 1* course teaches rigorous algebra fundamentals, and is the first self-paced course

[18] See, for instance, Muralidharan et al. (2019).

offered by the company. The course covers the first seven chapters of the *Prealgebra* textbook—Arithmetic, Exponents, Number Theory, Fractions, Equations and Inequalities, Decimals, and Ratios, Conversions and Rates—each consisting of several topics. In each chapter, the self-paced version supplements the chapter's readings with video lectures, sample problems, and additional resources. There are 366 distinct elements of the course, and we received data from AoPS on students' engagement with each of them.

The course and accompanying textbook are entirely in English. While all students in Tamil Nadu take classes to learn English, it is not their native language. To reduce potential language issues, we subtitled all video lectures in Tamil, but we were not able to translate any of the written parts of course including all quizzes and challenge problems. In practice, many students additionally used an online translator or the WhatsApp groups to resolve any language issues.

Two course components—Alcumus quizzes and challenge problems—allow us to assess student learning. The first, Alcumus, presents questions one at a time. The system adjusts the difficulty level in real time so that all students have roughly a 70% chance of answering a question correctly. Students can either "pass" or "master" each quiz by working up to and correctly answering questions of a certain difficulty level. The number of correct questions required to pass or master the topic varies; if a student starts with very challenging questions, it may take as few as two right answers to pass, but if the student makes mistakes, it may take many more. The majority of our students appear to move on once Alcumus tells them they have passed, but our most-engaged students persisted to gain mastery. While these problems are difficult, the adaptive, game-like system of Alcumus provides an environment that is not only educational, but fun for many students. The Prealgebra 1 course contains 36 such quizzes.

In contrast, the course's 144 challenge problems are non-adaptive. These problems are intended to challenge top students, and go far beyond direct applications of the material. While many students struggle with these problems, they can make as many attempts as they like to try to solve them and are told if their answers are correct. Students who solve these problems demonstrate a deeper mastery of the material.

From these data, we construct three measures of course engagement/performance on a single dimension. *Course Engagement* is the fraction of the 366 course elements that a student interacted with.[19] *Alcumus Passing* is the fraction of the Alcumus quizzes that a student passes or mastered. *Challenge Solved* is the fraction of challenge problems a student answers correctly. We also construct two indices reflecting performance on multiple dimensions. *Successful Completion* is an indicator variable for whether a student clicked on at least 80% of the course components, passed at least 90% of the Alcumus quizzes, and solved at least 60% of the challenge problems that they viewed. Many of our analyses focus on this easy-to-understand metric. We also define a

[19] Our results can be benchmarked to the average completion rate in AoPS' regular self-paced classes of about 70%.

continuous measure, *AoPS Index*, on a 0–1 scale that is a weighted average of several performance measures.[20]

3 Program Effectiveness: Student Engagement and Learning

The most important finding of our experiment is that our student identification and support procedures were sufficient to find a substantial number of students who engaged with and learned the course material. This finding is particularly remarkable given the challenges of our setting, including the gap between the difficulty of the material and the students' starting level of mathematics knowledge.

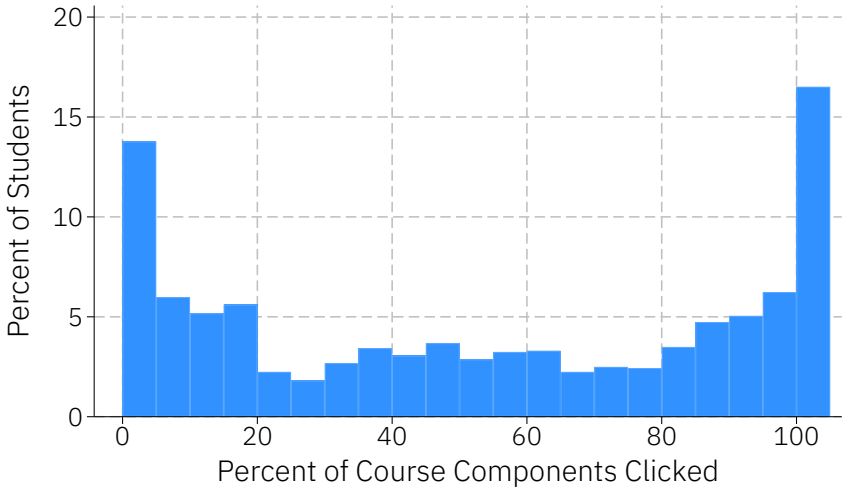


Figure 1: Histogram of course engagement of treatment students, as measured by the number of course components clicked.

Figure 1 presents data on course engagement as measured by the fraction of the 366 course components clicked on by each student. The distribution is bimodal. About 17% of students clicked on all 366 components of the course. Another 11% clicked on 90–99% of the components, and 8% clicked on 80–89%. There is also an equally large left tail, as 29% of students engaged with less than 20% of the course. The right tail is more important for our purposes, though, as our goal is to identify a subset of students who engaged with this course and might potentially be helped by a program of additional online courses. In total, 578 students (36% of those invited to participate) interacted with at least 80% of the course.

Extensive usage of the AoPS course does not necessarily coincide with extensive learning. Some students could have clicked all the links without ever learning the material or solving any prob-

[20] The index places weight one-third on the fraction of items clicked, one-third on the fraction of problems answered correctly (with weight one-half on challenge problems viewed and one-half on Alcumus problems viewed), one-sixth on the fraction of items clicked by June 1 (about four months into the program), and one-sixth on the number of problems completed (with weight one-half on challenge problems and weight one-half on Alcumus defined as quizzes passed).

lems.[21] Our primary source of information on whether students learned the course material is the Alcumus quizzes. Figure 2 presents a histogram showing the fraction of the quizzes that each student passed.[22] The fraction of students who passed all of the quizzes is 22%, which is even higher than the fraction who viewed all course components. We infer that many students not only engaged with the course but learned the material. The majority of the students who passed 100% of the Alcumus quizzes usually persisted beyond the passing mark and tried to master the material as well: the majority of those who passed all 36 quizzes mastered more than 90% of them.

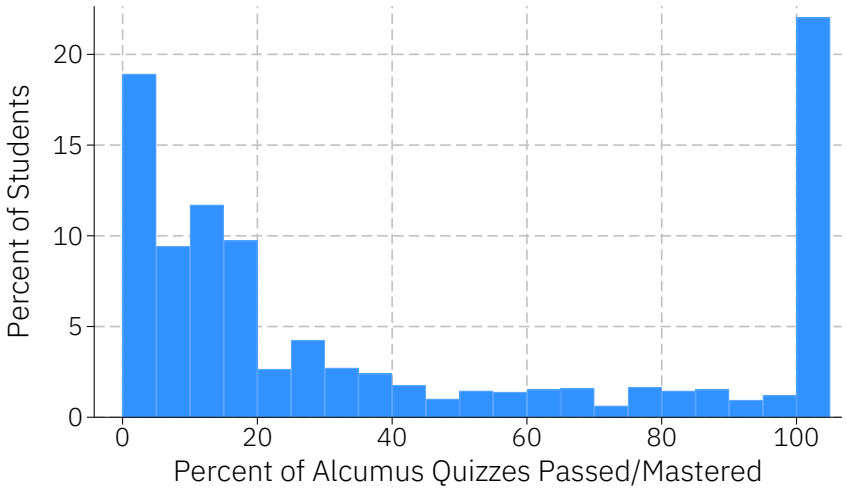


Figure 2: Histogram of course learning by treatment students, as measured by the percentage of Alcumus quizzes passed or mastered.

There is also less encouraging news in the figure: many students appear to have engaged in the course without having learned the content. The median student viewed 55% of the course components, yet passed just 20% of the Alcumus quizzes.

Even in the upper tail, the fraction who passed 80–100% of the quizzes is well below the fraction who viewed 80–100% of the course components: 27% vs. 37%. Some of this phenomenon is due to students who simply did not take the Alcumus quizzes—recall that the program was entirely voluntary. However, most students who viewed 80% or more of the course without passing 80% of the Alcumus quizzes did put in substantial effort: in this group, students viewed a median of 570 Alcumus problems. To make this concrete, two students viewed exactly 570 Alcumus problems. One got 62.5% of the questions correct, passing just 12 of the 36 quizzes; the other got 45% of the questions correct, passing just 9 quizzes. Each stuck with the course to the end: one clicked on 329 of the 330 non-Alcumus components and the other clicked on 328.

Indeed, the course data contain many stories of remarkable grit among students who found the

[21] Conversely, course components may also under-record a student's engagement if, for example, they watched videos together with another student while logged into that student's account.

[22] Students must pass quizzes before they master them, so all students who have mastered have also passed.

course very difficult. One student never passed a particular quiz despite having given 145 correct answers on it: Alcumus never passed them because they took a total of 372 questions to obtain the 145 correct answers. The overall leader in Alcumus problems attempted tried to answer 8,098. We infer that this student viewed so many problems because they aspired to master each of the quizzes despite the problems being hard for them—they got just 3,693 of the 8,098 problems correct. In the end, that student mastered 23 of the Alcumus quizzes and settled for passing the other 13.

The Alcumus problems also allow us to examine the relationship between student engagement and problem-solving ability. Figure 3 contains histograms of the percent of Alcumus quizzes passed or mastered, split by students' tercile of engagement.[23] The vast majority of students in the top tercile of engagement—65% of them—passed all Alcumus quizzes. Just two students who passed all Alcumus quizzes were in the middle tercile of engagement, and none were in the the lowest tercile. Importantly, the histogram on the left suggests that the difficulty of the course may be one factor that leads some students to make little progress: 65% of students in the lowest tercile of engagement passed exactly 0 Alcumus quizzes.

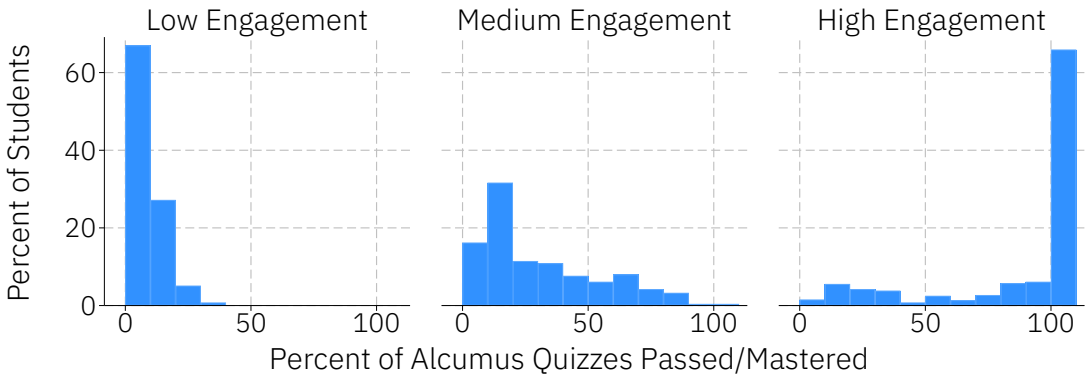


Figure 3: Histogram of course learning by treatment students, split by tercile of course engagement.

This conclusion is corroborated by Figure 4, which presents the fraction of Alcumus questions that students answered correctly for students with different levels of engagement. It provides a nice illustration of the adaptive technology: most students in all three terciles answer about 70 percent of their Alcumus questions correctly. Notably, the high-engagement tercile has an upper tail of very high performers who get 90%+ correct, while the low-engagement tercile has a set of low performers who get fewer than 10% correct. Perhaps these students gave up before Alcumus could adapt to their level. It's also possible that their baseline knowledge level is too far behind to handle even Alcumus' easiest problems.

[23] Again, we measure engagement using the number of course components clicked. Students in the lowest tercile interacted with 0–109 components (0–30%); the middle tercile, 110–310 components (30–85%); and the highest tercile, 311 to 366 components (85–100%).

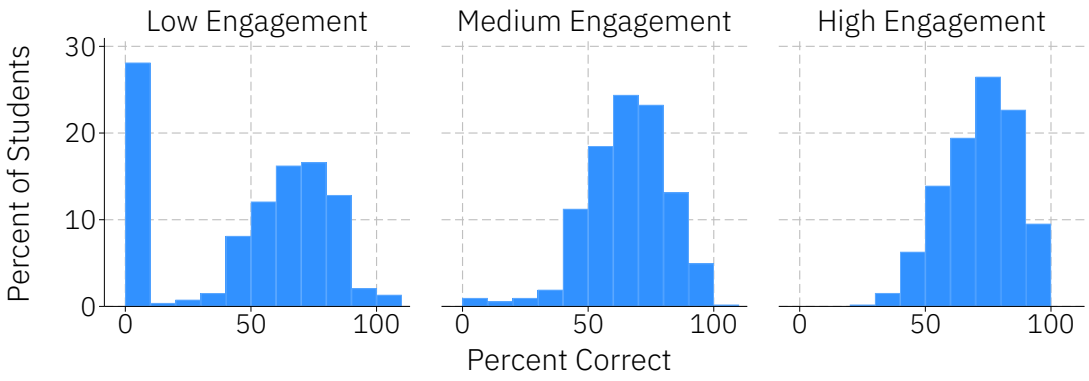


Figure 4: Histogram of correct answer rate in Alcumus, split by tercile of course engagement.

4 Program Design: Student Selection

One of our main objectives was to investigate the relative effectiveness of different criteria for choosing students for the program. To this end, we designed a set of subtreatments that use different combinations of criteria to choose students for the RCT. We randomized those subtreatments at the school level, so we can compare mean outcomes across subtreatments for a clean test of effectiveness. Note that we have measures of all criteria for all students, so we can also present regression-based results on the predictive value of each criterion.

Identifying students who can engage with and benefit from an enrichment program such as ours is a crucial piece of program design, yet it has been understudied in the economics of education literature. We think that our results fill an important gap.

As mentioned in Section 2, we used eight different methods to select students for the program, each one being some combination of four components: *Admin*, *Booklet*, *Quiz*, and *Grit*. Our initial interest is in which combinations of these components best predict engagement with and success in the AoPS course. We will also, of course, be interested in their ability to predict longer-run benefits from our interventions in later analyses.

In Table 4 we list five measures of course learning and engagement across the top and the eight combinations we used for selection down the left side. Each column of the table provides a comparison of means across selection methods. Columns (1) and (2) use measures that reflect both students' overall engagement and learning as described in Section 2.3. Column (3) is a measure of the number of course components interacted with, while columns (4) and (5) measure students' learning using their rates of passing Alcumus quizzes and correctly answering challenge problems, respectively.

Selection methods that use *Booklet* perform well. Students selected using *Admin+Booklet* were most successful according to every metric except for *Successful Completion*, where students selected using *Booklet* alone performed the best. Students selected using only *Admin* fare far worse on every measure than students selected using any other methodology: just 12% of these students successfully completed the course. One way to view these results is that combining two criteria meant to measure different attributes that might contribute to success—past administrative test scores to gauge ability and time spent on enrichment material to gauge enthusiasm—performs well in selecting students for a program like ours. Do recall, however, that all selection criteria also used the administrative test scores indirectly: the first-stage selection of 9,709 was primarily based on their 6th-grade test scores. These eight selection methods in this table were used to choose 2,490 students for our RCT from a population of students who mostly have at least pretty high administrative test scores.

Table 4: Statistics on course engagement and learning for treatment students selected using different methods.

	(1) Successful Completion	(2) AoPS Index	(3) Course Engagement	(4) Alcumus Passing	(5) Challenge Solved	Num. students	Num. schools
Admin+Booklet+Quiz	0.211	0.624	0.576	0.422	0.331	495	178
Admin+Booklet	0.256	0.652	0.585	0.454	0.372	224	79
Admin+Quiz	0.177	0.557	0.508	0.364	0.280	556	197
Booklet+Quiz	0.206	0.591	0.532	0.392	0.322	269	78
Booklet	0.283	0.620	0.571	0.450	0.358	183	78
Quiz	0.227	0.634	0.578	0.436	0.345	200	76
Grit	0.235	0.589	0.519	0.431	0.304	245	74
Admin	0.115	0.502	0.442	0.324	0.227	318	109

Table reports means for metrics of completion, learning, and engagement outcomes for students selected using each criterion.

The second way in which we examine selection is by regressing our five different measures of course performance in turn on student characteristics known at the time of their selection. Specifically, we use *Admin*, *Booklet*, *Quiz*, and *Grit*, all normalized to mean 0 and variance 1, as well as if a student was identified by their teacher as either a "top" student or a student who would benefit from the challenge of this course.[24] As the various selection methods selected students solely on the basis of right-hand-side variables in these regressions, they provide a valid way to investigate how one could have best predicted which students would be successful.

Note that as all continuous variables has been normalized, the coefficient estimates can be compared across measures as the effect of a one-standard deviation increase in the relevant variable. Results from these regressions are reported in Table 5.

[24]These last two regressors are described in further detail in Appendix A. We also include demographics such as gender, medium of school instruction, and whether they are in a scheduled caste or tribe; indicators for subtreatments they were assigned to; an indicator for if they cheated on the quiz; and indicators for missing quiz, grit, and administrative

Table 5: Regression of course engagement and learning on characteristics used for treatment selection.

	(1) Successful Completion	(2) AoPS Index	(3) Course Engagement	(4) Alcumus Passing	(5) Challenge Solved
Admin (z)	0.010	0.053***	0.050***	0.044***	0.045***
	(0.016)	(0.015)	(0.015)	(0.016)	(0.013)
Booklet (z)	0.069***	0.089***	0.089***	0.080***	0.087***
	(0.011)	(0.010)	(0.010)	(0.011)	(0.009)
Quiz (z)	0.015	0.026**	0.022*	0.020*	0.027***
	(0.012)	(0.011)	(0.011)	(0.011)	(0.010)
Grit (z)	0.015	0.016*	0.014	0.018*	0.013
	(0.010)	(0.009)	(0.009)	(0.009)	(0.008)
Teachers' Top	0.003	0.027	0.025	0.020	0.017
	(0.023)	(0.020)	(0.021)	(0.022)	(0.019)
Teachers' Challenge	-0.066**	-0.023	-0.018	-0.031	-0.030
	(0.031)	(0.029)	(0.029)	(0.031)	(0.025)
N	1,605	1,605	1,605	1,605	1,605
Mean	0.203	0.590	0.534	0.400	0.310

Standard errors in parentheses, and clustered at the school level. Table reports selected coefficients from regressing measures of completion, engagement, and learning on students' baseline academic characteristics and demographic and subtreatment controls.

$^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

These results are broadly in line with the results above from the selection randomization—*Booklet* is highly predictive for all measures of course performance and *Admin* is also an important predictor for most of them. The point estimates suggest fairly large effects: for instance, a student who is one standard deviation above average in *Booklet* is 6.9 percentage points more likely to successfully complete the course, a large effect relative to the mean completion probability of 20.3%. In contrast with the positive point estimates for the rest of the regressors, *Teachers' Challenge* is negatively associated with all measures of performance, but most are not significant.

5 Program Design: Support and Incentives

Helping talented students reach their potential can require much more than just providing them with access to online resources. In this section we present results on the tutoring, incentive, and computer lab access subtreatments (described in Section 2.2.2), which explore support components. Our most notable finding is that the online courses work quite well without any tutoring support.

score. We suppress those estimated coefficients in the table.

5.1 Tutoring

A number of previous papers have found that online education programs on their own may have limited impact. Tutoring is costly and challenging to implement at scale, but our prior belief was that tutoring could be a very valuable complement to online courses, and perhaps even necessary to make them effective our setting. As described earlier, we randomized students to either receive group tutoring, individual tutoring, or no tutoring.

The top portion of Table 6 presents statistics on *Successful Completion* for students in different tutoring treatments. We find that students *without* tutors did substantially better than the students in either tutoring arm: 28.5% of students without tutoring successfully completed course, while only 17.4% of students with individual tutoring and 10.5% of students with group tutoring did so. To assess the significance of the differences we regress *Successful Completion* on subtreatment dummies, baseline ability controls, and demographics.[25] The estimates, presented in the right column the table, show that the differences are highly significant. From a practical perspective this finding can be seen as very good news—the program is much cheaper and easier to scale up if tutors are not needed, and the results here indicate that the program without tutors works even better than the overall descriptive statistics indicated!

What explains the lack of tutoring effects in this study, especially when such support has been found to be valuable in other settings? It could be that the prevailing wisdom about the value of tutoring support is incorrect. However, it seems more likely that there is something special about our environment and intervention that weakens the effect of tutoring. One obvious possibility is that our tutors simply were not good. We do not think this is the case; after seeing no tutoring benefits in an earlier pilot, we invested heavily into tutor recruitment, training, and monitoring for our RCT. These tutors came from elite technical universities, worked diligently, were highly enthusiastic, and taught well.[26] Another feature that could play a role is that many students failed to complete the course. For students in group tutoring, students may have slowed their progress to match the progress of other students in their group. A third explanation is that our course itself may have been already highly engaging (perhaps to the extent of being a beneficial online addiction), and that time spent with tutors decreased time on the platform, and thus decreased their likelihood of getting "hooked." Future phases of our study may provide opportunities to help disentangle the multiple potential explanations.

[25] The regressions include five ability controls (end-of-6th-grade math scores, booklet engagement, exam score, grit, and teacher recommendation) three demographics (gender, medium of school instruction, scheduled caste/tribe membership). Standard errors are clustered at the school level. All results are very similar in regressions without controls.

[26] In discussions with tutors after the program, at least two tutors mentioned that being a tutor for our program led them to change their career paths to one in education.

Table 6: Course performance by subtreatment.

Subtreatment	Mean *Successful Completion*	Regression Estimate
Group Tutoring	0.105	-0.178***
		(0.023)
Individual Tutoring	0.174	-0.107***
		(0.036)
No Tutoring	0.285	—
Computer Lab Access	0.218	0.004
		(0.030)
No Access	0.197	—
Enforced Incentives	0.188	0.031
		(0.026)
Unenforced Incentives	0.240	0.052*
		(0.030)
No Incentives	0.185	—
N		1,605
Mean		0.203

Standard errors in parentheses, and clustered at the school level. Table reports the mean *Successful Completion* rate for various subtreatments groups and selected coefficients from regressing course completion on subtreatment indicators and ability and demographic controls.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5.2 Computer Labs

In this subtreatment, we provided students with after-school access to their school's computer labs; because this access was not always possible, estimates here should be interpreted as intent-to-treat estimates.[27]

Results are presented in the middle panel of Table 6. The students chosen for the computer lab subtreatment were slightly more likely to successfully complete the course, but the difference is not significant. Even so, we believe that the computer lab treatment was beneficial to students, but that benefit is not apparent because of unfortunate timing. As discussed above, we launched the program in late January, but computer lab access only began in late February or early March; students thus were only able to access the labs for about one month before access was lost due to

[27] Of students in the computer lab treatment group, 71.4% of them were able to access their school's Hi Tech Lab, 24.5% used an alternative computer lab within their school, and 4.1% of students were not able to use a computer lab.

final exams (early April) and the summer vacation (mid-April through early June). Figure 5 shows student engagement, as measured by number of course components clicked, for both the computer lab treatment and control groups. The groups were making quite similar progress up to the two-week period in which we started to roll out the subtreatment, after which the treatment group progressed at approximately twice the rate of the control group over the next month. Over the next two months, while treatment was paused, the gap between treatment and control faded. We conjecture that the computer lab treatment would have been substantially more effective if the course had been offered entirely during the school year with uninterrupted lab access.

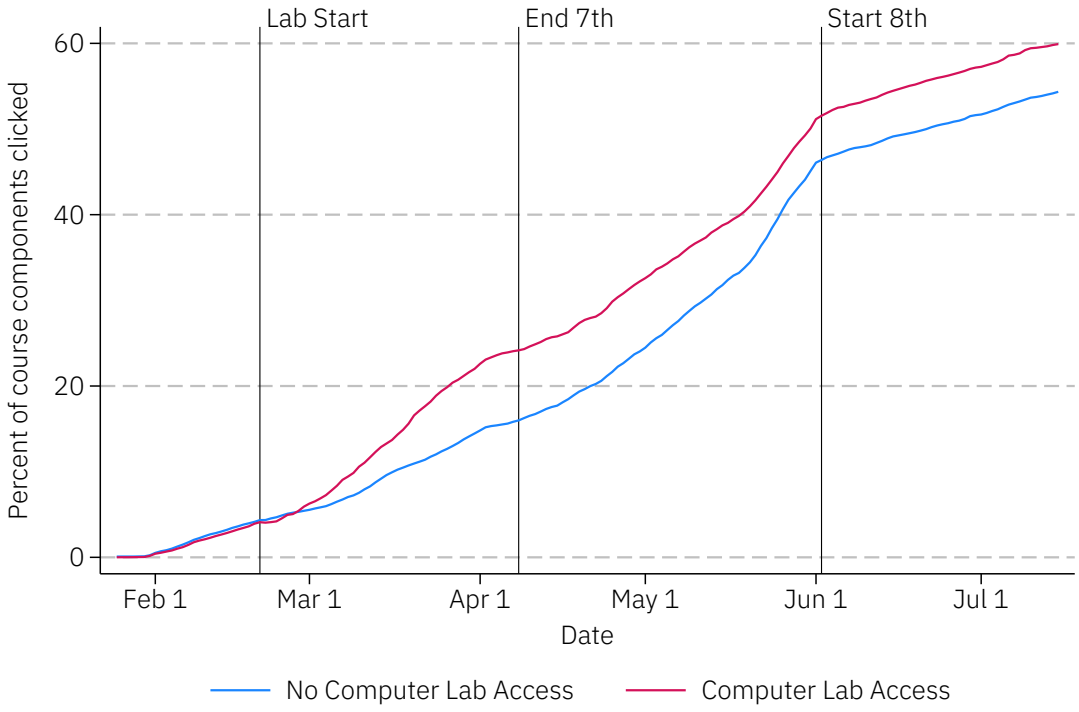


Figure 5: Progress of the computer lab treatment and control groups over time.

5.3 Incentives

A third dimension on which we varied the program was in the treatment of students who were not making sufficient progress in the course. Two subtreatment groups received messages warning them that tablets would be taken away if they were not making sufficient progress, though this warning was enforced for only one of the groups. The remaining third received no such messages. Progress for these different groups is described in the bottom panel of Table 6. Groups that received the warning methods were somewhat more engaged than the group that did not, but differences are not very large and again fall short of statistical significance.

In part, the failure of the warning messages to be effective may reflect that our messages did not

become highly specific until late in the summer vacation,[28] and no tablets were actually taken away until the final month of the program. Even if the incentive effects are small, we think it is advisable to include enforced incentives in the program. The data give no reason to think that the incentive effect is negative, and if students are using their tablets to watch YouTube video rather than to work on AoPS we would conjecture that it is better for them to have the tablet taken away.

6 Program Effectiveness: Heterogeneity

Here we present evidence on how course engagement and achievement differed across demographic groups. We find that the expected socioeconomic gaps are present in our population, with one notable exception: girls outperform boys.

Table 7 reports mean *Successful Completion* rates for students in various demographic groups and regression coefficients that can be used to assess the significance of the differences.[29]

We look first at the gender breakdown and find that girls were 25% more likely than boys to achieve the *Successful Completion* benchmark based on a comparison of means. When we control for observed characteristics, the estimated coefficient on gender is not statistically significant, but the point estimate indicates that girls were still about 15% more likely to successfully complete the course.[30] As mentioned earlier, about two-thirds of the students selected for our program are girls, despite the fact that our selection was gender-blind. This itself is noteworthy given the gender gaps present in elite Indian education: in the mid-2010s, only 8% of students at the Indian Institutes of Technology were women.[31]

Our experiment also highlights the large socioeconomic gaps in STEM achievement present in India. Recall that our initial set of schools considered for potential participation in our program reflected in part a desire to see if we could identify high-achieving students from disadvantaged groups: 38.1% of the students at the 3,494 schools in our initial consideration group were members of a scheduled caste or scheduled tribe, groups that are particularly disadvantaged in an environment where *all* students are disadvantaged to some extent. We did not use any form of demographic-based affirmative action in selecting students for the RCT, and 23.6% of our result-

[28] Taking away tablets during the summer vacation would have incurred substantial costs, which we aimed to avoid.

[29] Results are qualitatively similar if other measures of course engagement/performance are used as the dependent variable in place of *Successful Completion*. Following our pre-analysis plan, which did not specify including fathers' education and employment in our regression, we omit those variables from the regression, but we report the mean completion rates to provide a comparison of how students' socioeconomic status impacts their course engagement.

[30] In unreported regressions including interactions of the male indicator with the various subtreatment dummies, we did not find evidence that the results are driven by any treatment being particularly good or bad for boys.

[31] See Madras (2020). IIT Madras has since started an affirmative action program for female students, and about 20% of current undergraduate students are women. The reverse gender gap, however, is consistent with Tamil Nadu's results from the PISA 2009 Plus, which reported that girls were twice as likely as boys to score in the Level 3 and above ranges: 4.6% vs. 2.3% (Walker, 2011). The PISA 2009 Plus had 6 proficiency levels, ranging from level 1 ("routine mathematical tasks and procedures" to Level 6 ("advanced mathematical skills and reasoning") Students at Level 2 display a basic competency in mathematics.

Table 7: Heterogeneity in successful completion rates.

	Mean *Successful Completion*	Regression Estimate
Female	0.218	0.026
		(0.022)
Male	0.172	—
Scheduled Caste/Tribe	0.148	-0.046**
		(0.023)
Not Scheduled Caste/Tribe	0.220	—
Non-English Medium School	0.182	-0.041*
		(0.022)
English Medium School	0.230	—
Father's Employment Type		
Daily Wage	0.181	
Private	0.252	
Government	0.318	
Father's Education Level		
5th Grade and below	0.192	
8th Grade	0.144	
10th Grade	0.173	
12th Grade and above	0.283	
N		1,605
Mean		0.203

Standard errors in parentheses, and clustered at the school level. Table reports the mean *Successful Completion* rate for various demographic groups and selected coefficients from regressing course completion on demographic variables and subtreatment and ability controls.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

ing sample belongs to a scheduled caste or tribe. These students made somewhat less progress in the course than the other students, both on a raw basis and controlling for baseline characteristics. The regression coefficient indicates that they were about 4.5 percentage points less likely to successfully complete the course. While the gap results in only 17% of our successful completion group being from a scheduled caste or tribe, an optimistic take is that we have identified 57 students from a scheduled caste or tribe who met our successful completion benchmark. These students did quite well, correctly answering 86% of the challenge problems they viewed and mas-

tering 75% of the Alcumus quizzes.

Just under half of our students attend English-medium schools, while the remainder of students take most of their courses in Tamil. At the start of our experiment, both the Tamil Nadu School Education Department and our research team were concerned about whether Tamil-medium students would find difficulty with a course entirely in English.[32] Tamil-medium students had slightly lower rates of course completion than their English-medium counterparts (18.2% vs. 23.0%), though this difference is not statistically significant after controlling for subtreatments, baseline ability and other demographics. The difference in *Successful Completion* rates stems primarily from differences in the rates at which students passed the Alcumus quizzes: *Course Engagement* is similar for the English- and Tamil-medium students. The difference in *Alcumus Passing* could be due either to Tamil-medium students having not learned the material as well or to their having difficulty correctly parsing some questions. One note for future analysis is that some tutors reported that their Tamil-medium students improved their English skills over the course of the program; we plan to examine this potential side-benefit of the program once we receive students' scores in 8th-grade English classes.

Student outcomes are also (mostly) aligned with parental employment and education in the expected direction. Successful completion rates are 17% for students whose fathers have a 10th grade education (the modal value) vs. 28% for students whose parents have a 12th grade or college education. This suggests that parental support is still playing a role in student success in our course. The majority of our students' fathers are "daily wage" workers, which indicates that they are paid when they work and is often associated with working in the informal sector. Students whose fathers have more stable private-sector or government jobs (14% and 3% of the RCT sample) are again substantially more likely to successfully complete the course.

7 Program Effectiveness: Broader Learning

In this section we examine whether treated students broadly improved their math knowledge and problem-solving skills.

As discussed in Section 3, student performance on Alcumus quizzes shows that students who completed *Prealgebra 1* mastered many topics, often after a great deal of effort. Given this, it did not seem particularly valuable to compare treatment and control students on a final exam testing only material from the treatment course. Instead, we planned to look for evidence of broader learning in two ways: first, by administering a challenging exam assessing the ability to solve very difficult problems on topics from the school curriculum; second, by examining scores on exams that are administered in the schools to all Tamil Nadu students.

[32] As discussed in Section 2.3, we added Tamil subtitles to the course videos, but the videos themselves, the written lessons, and all problems were only in English.

We designed our challenging exam based on the standard Tamil Nadu math curriculum in 6th–8th grade. Most questions were focused on prealgebra, but several questions used concepts from algebra, geometry, and statistics. Given the prevalence of cheating on the quiz we used in our initial selection, we worked hard to prevent and detect cheating: hiring proctors, not using multiple choice, altering question orders, and changing numbers in several questions to change the correct answers.

Unfortunately, our test was much more challenging than we had intended. The test started with one simple word problem that 89% of students answered correctly. However, our difficulty quickly ramped up to levels that almost all students could not handle, with a mean score of just 2.23 out of 20.[33] As a result, the standard deviation of test scores (1.73) was quite large relative to the mean, limiting our ability to detect moderate increases in the percentage of problems students can solve. One redeeming feature of our too-difficult test is that because students incorrectly answered so many non-multiple choice questions, it was easy to detect if students copied; we label 12% of students as having cheated.

Table 8 presents a comparison of scores in the intended treatment and control populations.[34] The left panel presents results for all students in the treatment and control populations who took the test.[35] The difference between the mean treatment and control scores is 0.11, a difference of approximately 5% of the control mean. We further analyze differences and assess significance via a regression of scores on the challenging exam on baseline ability and demographic controls and the treatment dummy. The difference in scores fails to be significant; given the standard error on the estimated treatment effect, we are not able to identify increases of less than 10% in the number of problems solved.

The right panel repeats the same analysis, but excluding students who cheated on the exam; the estimated difference increases to 0.20. One can think of the magnitude of the effect as a 9% increase in the number of problems solved, or as a 16% increase in the number of challenging problems solved. The estimated treatment effect on this subsample is statistically significant. We find it plausible that there might have been more effective cheaters in the control population because teachers in control schools might have been less willing to help the proctors prevent cheating, as many teachers/headmasters were disappointed that their schools were not selected. However, the significance of the score difference depends on whether one is willing to assume that the cheaters are sufficiently similarly selected in the treatment and control populations, and thus that a comparison of the non-cheating students is appropriate.

[33] The percent correct on the next seven questions were 38, 20, 20, 13, 11, 8, and 4; the remaining twelve questions were each correctly answered by less than 3% of students.

[34] Recall that these should be thought of as intent-to-treat estimates because a few "treatment" students did not experience the program and a few "control" students did.

[35] We were unable to test 103 of the 2,490 students. Most were due to student absences, but a few students were present and declined to take the test. Missing scores are somewhat more common in the control population: 4.9% vs. 3.7%.

Table 8: Scores on challenging exam by treatment group.

(a) All

	Num. Students	Mean	Regression
Treatment	1,545	2.28	0.13
			(0.10)
Control	842	2.17	—

(b) Non-cheaters

	Num. Students	Mean	Regression
Treatment	1,342	2.29	0.20**
			(0.09)
Control	747	2.11	—

Standard errors in parentheses, and clustered at the school level. Table reports the mean scores on our challenging exam by treatment group and regression estimates of the treatment effect with ability and demographic controls. Panel (a) includes all students who took the challenging exam. Panel (b) restricts the sample to students who were not identified as having cheated.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Our regressions also included unreported controls for demographics and for the various measures of ability and potential-for-engagement that we used to select students for the RCT. The demographic effects are mostly what one would expect and similar to the demographic effects on course completion: scheduled caste/tribe students and students attenting Tamil-medium schools score lower. One contrasting result is that on gender: the boys had higher average scores than the girls. This is true both in the treatment and in the control.

We have test scores for two additional math tests administered by the schools as part of their ordinary practice. Students took their 7th grade annual exams approximately two months into our program. And they took the first midterm exam of their 8th grade year around the time that our program ended. The end-of-year exam is considered important and reported by schools to the school education department. The first midterm exam is less focal: it occurs about one-sixth of the way through the school year and is not reported to the school education department. We collected these scores by contacting teachers directly. Table 9 reports mean treatment and control scores and estimated treatment effects on these exams. We did not anticipate that any effect would be visible on the seventh-grade exam–the average student had only clicked on about 20% of course components at that point–and indeed we find no effect. We had hoped that we might see an effect on the eighth grade exam despite the limited content overlap, but again the estimated treatment effect is not significant.

An interesting pattern in the test score data is that the gap between the successful completers and other treatment students appears to be widening over time. Table 11 reports mean scores on the grade 6, 7, and 8 exams, comparing treatment students who did and did not achieve *Successful Completion*. The 6th grade scores are very similar, reflecting that 6th grade scores are not a very strong predictor of subsequent engagement (within the set of high-scoring students we selected). On the grade 7 exam—two months into the course—the students who will go on to be successful completers score better. This finding may reflect that they have higher ability, are more interested

Table 9: Scores on seventh- and eighth-grade math exams by treatment group.

(a) Grade 7 End of Year

	Num. Students	Mean	Regression
Treatment	1,377	84.3	-0.05
			(0.67)
Control	785	84.8	—

(b) Grade 8 First Midterm

	Num. Students	Mean	Regression
Treatment	1,420	68.6	1.01
			(1.45)
Control	795	68.1	—

Standard errors in parentheses, and clustered at the school level. Table reports the mean scores on two school-administered math exams (grade 7 end-of-year and grade 8 first midterm) by treatment group and regression estimates of the treatment effect with ability and demographic controls.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

in learning math, or are improving their broader math skills in the course. By the grade 8 exam—around when the program finished—the gap has further widened. This result strongly suggests that the gap is not just about differences in unobserved ability. Instead, it is evidence that the successful completers learned from the work they did over the final three months or that differences in motivation are contributing to *Successful Completion* and developing performance differences.

Table 11: Test scores by *Successful Completion*

	Sample Mean Scores		
	Grade 6 End of Year	Grade 7 End of Year	Grade 8 Midterm
Successful Completers	89.8	88.0	75.0
	(0.56)	(0.61)	(1.19)
Other Treatment Students	89.2	83.4	66.9
	(0.27)	(0.37)	(0.63)

Standard errors for sample mean estimates in parentheses. Table reports the mean scores for treatment students who did and did not achieve *Successful Completion* of course.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

8 Side Effects: Tablet Use and Other Academic Subjects

In this section we explore potential unintended consequences to the program. We note that students made substantial non-academic use of their tablets, but do not find that this usage had substantial adverse effects.

Students needed access to a variety of applications to participate in our treatment. Our course is accessed via a web browser; embedded videos are hosted on YouTube; and contact with students and parents (including tutoring sessions) was primarily through WhatsApp.[36] In the era of short-

[36] We also hoped that building community by organizing WhatsApp groups would boost student progress.

form video and social media, it is plausible that providing tablet access to teenagers who otherwise have little internet access could have nontrivial adverse effects.[37]

Our experimental design—reclaiming tablets from students when we visited schools to give our challenging exam—provides us with an opportunity to examine how students used tablets and to look for adverse effects.[38] Table 12 provides statistics on amount of time students spent using apps in the twelve app categories with the highest usage. Time spent on the AoPS course would appear as time spent in a browser.

Table 12: Daily tablet use by app category.

App Type	Most-Used App	Mean	SD
Video Players & Editors	YouTube	2.47	2.27
Communication	WhatsApp Messenger	0.55	0.43
Tools	Google Play services	0.28	0.31
Browser	Google Chrome	0.28	0.25
Phone	Phone by Google	0.17	0.21
Games	Google Play Games	0.13	0.27
Productivity	Google Drive	0.12	0.16
Photography	Gallery	0.10	0.14
Travel & Local	Google Maps	0.07	0.12
Music & Audio	YouTube Music	0.06	0.12
Education	Tamil Nadu 8th Guide	0.03	0.12
Social	Instagram	0.03	0.16

Table contains data the most-used application, mean hours used per student, and standard deviation of per-student daily usage (in hours) for each of the 12 most-used categories of applications. Table represents usage patterns of 1,254 students.

The most striking and potentially concerning aspect of the time-use data is that the tablets were used to watch a very large number of YouTube videos—about 2.5 hours per day on average. This is potentially concerning, but not necessarily so: students had the tablets for the entirety of their summer vacation, and some of the watching may have been done by family members. Students spent about 30 minutes per day on WhatsApp, Google Messages, Gmail, and other communication apps, and an additional 10 minutes per day using the Phone app. Some (perhaps surprisingly) positive news is that gaming and social media app use were quite low: just 8 and 2 minutes per day. Students could not have spent too much more time accessing games and social media via a browser, as browsers were used for just 17 minutes per day.

We use the 7th grade end-of-year exams and the first 8th grade math midterm to look for adverse

[37] TikTok has been banned in India since 2020; other short-form video platforms are prevalent in its place.

[38] We were able to recover usage data from just over three-quarters of the treatment students' tablets. Data could be unrecoverable for various reasons including if the student had changed the tablet password or wiped the data with a factory reset before returning the tablet. The students with missing usage data made substantially less progress on the course.

effects of time-wasting activities. Our students took their 7th grade end-of-year exams in mathematics, science, English, and Tamil about two months into the program. While students would not have gone through much of the course at that point, we could potentially see adverse effects if students watched YouTube videos when they should have been studying for final exams. The first four columns of Table 13 present estimated treatment effects from regressions with 7th-grade test scores as the dependent variable. The fifth column shows the same treatment effect on the 8th grade midterm math score we showed in the previous section. None of the estimated treatment effects are statistically significant.

Table 13: Intent-to-treat effects on seventh- and eighth-grade exam scores.

	7th-Grade End of Year Scores				8th Midterm
	(1) Math	(2) Science	(3) English	(4) Tamil	(5) Math
Treatment	-0.05	-0.09	-0.79	0.00	1.01
	(0.67)	(0.72)	(0.72)	(0.58)	(1.45)
Control Mean	84.8	86.3	85.6	89.9	68.1
Control SD	11.8	12.6	12.3	10.9	20.0

Standard errors in parentheses, and clustered at the school level. Table reports coefficients from regressing 7th-grade test scores on treatment status and ability and demographic controls.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

We can also use the student-level data to look for adverse effects in the cross-section of treatment students. We do so by regressing test scores and other outcomes on average daily app usage (in hours) for each of the 12 most-used app categories (listed in Table 12) and our usual set of ability and demographic controls. In this regression, we would expect test scores to be negatively associated with time spent on time-wasting apps for two reasons: first, selection into who uses the apps, and second, the negative causal effects of the apps. Hence, the regression coefficients should be upper bounds on the adverse causal effects.

Table 14 reports estimates from regressions with 7th and 8th grade test scores as the dependent variables. The first five rows present coefficients for the five most worrying categories of apps. There is no evidence whatsoever that the large amount of time spent on YouTube is having adverse effects. The most concerning negative result is that several of the associations with time spent gaming are significantly negative; spending an extra hour per day gaming is associated with scoring about 0.2 standard deviations lower on the 7th grade end-of-year exams.[39] Part of this is presumably a selection effect; even if the causal effect were this large, the aggregate effect on our treatment population would be modest given that our average student is spending just 8 minutes per day gaming. Spending time on the phone app also has significant negative associations with

[39] Note that this is relative to the standard deviation in our high-achieving pool; it is a smaller fraction of the standard deviation of the full distribution of test scores.

multiple scores.

Table 14: Relationship between student outcomes and app use.

	7th-Grade Scores				8th Midterm	Program Outcomes	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Math	Science	English	Tamil	Math	*Successful Completion*	Challenging Exam
Video Players	0.09	0.07	0.26	0.17	-0.09	-0.01	-0.03
	(0.20)	(0.20)	(0.20)	(0.14)	(0.35)	(0.01)	(0.03)
Communication	1.41	-2.78***	-0.95	0.84	0.89	-0.03	-0.02
	(0.88)	(0.91)	(0.98)	(0.73)	(1.66)	(0.04)	(0.13)
Phone	-3.39*	-2.64	-2.66	-4.36***	-9.40***	-0.04	-0.44
	(1.92)	(1.78)	(2.10)	(1.30)	(3.40)	(0.07)	(0.27)
Games	-1.76	-2.89**	-2.56**	-2.61***	2.17	-0.03	0.11
	(1.18)	(1.20)	(1.16)	(0.99)	(1.99)	(0.06)	(0.18)
Social Media	2.20	-4.07**	1.29	-1.26	3.41	0.04	-0.24
	(1.43)	(2.00)	(1.53)	(2.21)	(2.68)	(0.06)	(0.21)
Browser	3.61**	4.60***	4.26***	3.01**	13.27***	0.52***	1.00***
	(1.45)	(1.61)	(1.59)	(1.22)	(3.09)	(0.06)	(0.27)
Education	2.09	7.40***	10.19***	3.70	12.28***	0.16	0.86
	(2.27)	(2.58)	(2.97)	(2.28)	(4.11)	(0.10)	(0.72)
N	1,083	1,083	1,083	1,083	1047	1,164	999
R^2	0.201	0.155	0.184	0.212	0.168	0.117	0.113
Control Mean	84.8	86.3	85.6	89.9	68.1	–	2.2
Control SD	11.8	12.6	12.3	10.9	20.0	–	1.7

Standard errors in parentheses, and clustered at the school level. Standard errors in parentheses, and clustered at the school level. Table reports selected coefficients from regressing 7th-grade test scores and program outcomes on time spent using various application categories with ability and demographic controls.
$^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

The final two rows of Table 14 present estimated associations with time spent on "good" apps: the web browser (through which students accessed AoPS) and educational apps (the most popular of which is a curriculum guide produced by the Tamil Nadu Education Department). Time spent in the browser is significantly positively associated with scores in all subjects. The fact that more browser time is associated with positive scores on all exams (including Tamil) suggests that part of the effect is a selection effect. The fact that the effect is larger on the 8th grade exam is again also consistent with students who are using the browser having gained more over the full course of the program than they did just in the first two months. Time spent on education apps is significantly positively associated with seventh-grade science and English exam scores and the eighth grade math midterm score.

Column (6) of Table 14 looks at associations between app use and successful completion of our course. Again, there is little to suggest that the fact that we did not prevent students from accessing

YouTube videos is having an adverse effect the likelihood that they will finish the course. The co-efficient on browser-time is enormous and tremendously significant. This has a large mechanical component—students cannot finish the course without spending a lot of time in the browser—but does suggest that browser time is an effective proxy for time spent on AoPS. Column (7) examines associations with scores on the challenging exam we administered at the end of the program. Time spent in the browser has a large and highly significant association with scores: spending 30 more minutes per day in the browser is associated with scoring roughly 25% higher on our challenging exam. Again, we cannot say to what extent this is a selection effect or a causal effect.

In summary, our students spent a great deal of time watching YouTube videos, but the data so far do not indicate that this is a major concern. Results on how little time was spent gaming and on social media are good news, and certainly, we are not seeing effects that are anything like those in Malamud and Pop-Eleches (2011). This could reflect that concerns about adverse effects of the internet are smaller for students drawn from the upper tail of the distribution. Nonetheless, in future phases of the project we plan to be more vigilant and to more quickly take tablets away from students who are not making good progress on the classes.

9 Conclusion

India has a severe missing talent problem. The 7th graders in government schools in our study population have many obstacles standing between them and STEM degrees at top universities. Their schools do not have gifted education programs. They have little access to internet resources. Their parents are not positioned to help them with STEM material. Many are girls in a society in which girls are also extremely underrepresented in STEM. Our students are well behind similarly talented US students in their math knowledge, and they are falling further behind. It is hard to imagine that they could overcome these obstacles other than with a sustained multi-year effort.

The RCT described in this paper is the first stage of such an effort. The design of this effort reflects several hypotheses: 7th graders are old enough to be able to engage independently with online courses; providing cellular-enabled tablets is possible at scale; offering an online class to a larger group is a good way to identify students who will engage with a multi-year program of such classes; an aggressive approach, starting with compelling courses designed for students with greater math knowledge, can work for students who are bright enough and have enough interest and grit.

From this perspective, we view our RCT as a success: we have identified a number of students who engaged with and did very well in a challenging course, and we identified a number of additional students who displayed a great deal of grit while struggling to adapt to the level of the course. We cannot yet know whether the students who succeeded in one course will stick with and learn from future courses, nor if that will be enough to open up STEM careers, but the results are promising.

One promising aspect is what we find on selection methods. Our methods included a mini-version

of the idea that offering one course would be a good way to find students who will take advantage of future courses— we gave students a booklet containing one textbook chapter and measured engagement with it. We find that engagement with the booklet is a powerful predictor of successful course completion: a one standard deviation increase in booklet engagement is associated with a 6.9 percentage point increase in the probability of successful course completion.

Another striking finding, certainly good news for the long run, is that it is not necessary to provide students with tutors to make it possible for them to succeed. In fact, these tutors had a negative effect on student progress. Instead, a simpler-to-scale intervention with low-touch messages and incentives encouraged student success.

A third interesting finding in this paper is also reassuring. It is a legitimate worry that allowing students access to addictive online content might mean that any benefits of online courses could be offset by negative impacts. Our experiment provided an unusual opportunity to get a detailed look at how our students used their tablets and how this use was associated with their school performance. We do find that students (or their families) spent a tremendous amount of time watching YouTube videos. But, we also find quite limited time spent gaming and on social media, and do not find adverse effects of YouTube watching on students' schoolwork. It may be that aspects of our environment—focus on talented students, few peers having social media accounts, cultural norms, many non-working mothers—make tablet abuse less problematic here than it would be elsewhere.

Taken together, the results on selection and support suggest that one might be able to implement a version of our program that worked even more efficiently. 20.3% of the students we invited to join our program successfully completed the online course. But the successful completion rate was 28.5% for students who were not assigned to tutors. And a selection method using a combination of administrative scores and booklet engagement should also bolster completion rates. The fact that tutors seem unnecessary should greatly facilitate scaling.

The results in this paper suggest multiple avenues for future research. It is striking that the non-tutoring version of the program worked so much better than the versions with tutors. It would be very interesting to learn more about what in our setting leads to this unusual result. The idea of trying out mini-versions of programs on a larger population also seems like one that could have many potential applications.

With regard to our main questions of interest, the most important topic for further study is, of course, whether a multi-year effort focused on students who seem well-matched to online coursework will be sufficient to allow a substantial number of disadvantaged students to pursue STEM careers. In this experiment, we have mostly relied on students passing Alcumus quizzes and solving challenge problems as evidence that students are gaining from the AoPS courses. In the long run, we will be much more concerned with whether students are moving forward on the path to STEM higher education and careers. There are many more obstacles to overcome. But we are

starting from a base where essentially no students in the population we study ever reach the upper levels of the Indian educational hierarchy, and we are keen to see if our program can reverse this.

References

Abdulkadiroğlu, Atila, Josh Angrist, and Parag Pathak (2014) "The Elite Illusion: Achievement Effects at Boston and New York Exam Schools," *Econometrica*, 82(1), 137–196.

Albo, Fernanda, Salome Aguilar Llanes, Bernardo Garcia, and Sebastian Guevara (2025) "Expanding Access to Tutoring: A Scalable Platform for Personalized Learning and Data-Driven Research," *Proceedings of the 2025 AIED 26th International Conference on Atifical Intelligence in Education*.

Angrist, Noam, Micheal Ainomugisha, Sai Pramod Bathena et al. (2023) "Building Resilient Education Systems: Evidence from Large-Scale Randomized Trials in Five Countries," *NBER Working Papers*, 10.3386/w31208.

Beuermann, Diether W., Julian Cristia, Santiago Cueto, Ofer Malamud, and Yyannu Cruz-Aguayo (2015) "One Laptop per Child at Home: Short-Term Impacts from a Randomized Experiment in Peru," *American Economic Journal: Applied Economics*, 7 (2), 53–80, 10.1257/app.20130267.

Bui, Sa, Steven Craig, and Scott Imberman (2014) "Is Gifted Education a Bright Idea? Assessing the Impacts of Gifted and Talented Program," *American Economic Journal - Economic Policy*, 6(3).

Card, David and Laura Giuliano (2024) "Targeted Acceleration in Middle School Math: Impacts on College Entry, Degree Completion, and STEM," *Unpublished*.

Carlana, Michela and Eliana La Ferrara (2021) "Apart but Connected: Online Tutoring and Student Outcomes during the COVID-19 Pandemic," *HKS Faculty Research Working Paper Series*, RWP21-001.

Cimpian, Joseph R. and Jo R. King (2024) "An Institution-Level Analysis of Gender Gaps in STEM Over Time," *Science*, 386 (6724), 853–856, 10.1126/science.adr0787.

Cohodes, Sarah (2020) "The Long-Run Impacts of Tracking High-Achieving Students: Evidence from Boston's Advanced Work Class," *American Economic Journal: Economic Policy*, 12(1).

Dobbie, Will and Roland G. Fryer (2014) "Exam High Schools and Academic Achievement: Evidence from New York City," *American Economic Journal: Applied Economics*, 6(3), 58–75.

Duckworth, Angela L., Christopher Peterson, Michael D. Matthews, and Dennis R. Kelly (2007) "Grit: Perseverance and Passion for Long-Term Goals.," *Journal of Personality and Social Psychology*, 92 (6), 1087–1101, 10.1037/0022-3514.92.6.1087.

Ellison, Glenn and Parag A. Pathak (2025) "Optimal School System and Curriculum Design: Theory and Evidence," Working Paper34091, National Bureau of Economic Research, 10.3386/w34091.

Ellison, Glenn and Ashley Swanson (2010) "The Gender Gap in Secondary School Mathematics at High Achievement Levels: Evidence from the American Mathematics Competitions," *Journal of Economic Perspectives*, 24 (2), 109–128, 10.1257/jep.24.2.109.

——— (2016) "Do Schools Matter for High Math Achievement? Evidence from the American Mathematics Competitions," *American Economic Review*, 106 (6), 1244–1277, 10.1257/aer.20140308.

Escueta, Maya, Andre Joshua Nickow, Philip Oreoploulos, and Vincent Quan (2020) "Upgrading Education with Technology: Insights from Experimental Research," *Journal of Economic Literature*, 58 (4), 897–996, 10.1257/jel.20191507.

Hoxby, Caroline and Christopher Avery (2012) "The Missing "One-Offs": The Hidden Supply of High-Achieving, Low Income Students'," Working Paper18586, National Bureau of Economic Research, 10.3386/w18586.

Hoyos, Carola (2019) "The Biggest Gender Divide Is in Mathematics," https://www.ft.com/content/8d9dba28-ae17-11e9-b3e2-4fdf846f48f5, September.

Koedinger, Kenneth R, John R Anderson, William H Hadley, and Mary A Mark (1997) "Intelligent Tutoring Goes To School in the Big City," *Carnegie Mellon University*, 10.1184/R1/6470153.v1.

Kraft, Matthew A. and Virginia S. Lovison (2024) "The Effect of Student-Tutor Ratios: Experimental Evidence from a Pilot Online Math Tutoring Program," 10.26300/87MA-E949.

Madras, IIT (2020) "IIT Madras Offers Unparalleled Academic Flexibility to Students," technical report.

Malamud, Ofer, Santiago Cueto, Julian Cristia, and Diether W. Beuermann (2019) "Do children benefit from internet access? Experimental evidence from Peru," *Journal of Development Economics*, 138, 41–56, 10.1016/j.jdeveco.2018.11.005.

Malamud, Ofer and Christian Pop-Eleches (2011) "Home Computer Use and the Development of Human Capital," *Quarterly Journal of Economics*, 126 (2), 987–1027, 10.1093/qje/qjr008.

Moreira, Diana (2019) "Success Spills Over: How Awards Affect Winners' and Peers' Performance in Brazil," 72.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian (2019) "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India," *American Economic Review*, 109 (4), 1426–1460, 10.1257/aer.20171112.

Saavedra, Santiago and Tom as Buitrago (2025) "Discovering and Supporting Hidden Talents through Math Olympiads in Colombia," *Working paper.*

Steegh, Anneke M., Tim N. Höffler, Melanie M. Keller, and Ilka Parchmann (2019) "Gender Differences in Mathematics and Science Competitions: A Systematic Review," *Journal of Research in Science Teaching*, 56 (10), 1431–1460, 10.1002/tea.21580.

Walker, Maurice (2011) "PISA 2009 Plus Results: Performance of 15-year-olds in reading, mathematics and science for 10 additional participants," technical report, Australian Council for Educational Research.

A Methodology for initial screen

Here, we describe the methodology used to select the 9,709 students who were eligible for selection into our RCT.

1. Beginning with our population of 112,630 7th-grade students students in 3,494 Tamil Nadu government schools, we first identified students we believed had a high probability of succeeding in the program. At each school, we identified the number of students that meet the following criteria:

   - A "very high achiever" is a student with a score of at least 90 on their 6th grade final exam, and a within-school z-score of over 2.
   - A "high achiever" is a student with a score of at least 80 on their 6th grade final exam, and a within-school z-score of over 1.5.

   We selected a school for inclusion if either

   - the school contains at least 1 very high achiever, or
   - the school contains at least 2 high achievers.

   This procedure allows us to target selection towards schools where we expect to find students that can benefit from our program, and selected 1,441 schools that enroll 73,533 7th-grade students in total.

2. At included schools, our research team selected the following students were offered the sample AoPS content:

   - Students with a raw score of at least 70, a z-score of over 1.2, and with $30z + \text{score} > 125$. (All criteria needed to be met to be chosen.) In doing so, we attempted to capture students with high achievement in mathematics..
   - For schools with more than 5 students selected using the above criteria, we also, with probability $\frac{1}{2}$, selected $\sim$15% more students randomly from those above the 75th percentile within their school. This allowed us to explore how selection would work for students lower in the achievement distribution.

   This procedure selected 7,272 students across the 1,441 schools.

3. As teachers may have more informative signals about student ability than that observed in the administrative data, the final set of students was chosen using the following procedure.

   (a) At each school, we described the study and sample content to the math teacher. We then ask them to identify the top 5% of students in mathematics in their school (minimum of 1) to receive the sample AoPS content.

(b) We then revealed the list of students described in step (2).

(c) Finally, we asked the teacher to identify the top 2.5% of students (minimum 1) who have "high potential" – students that may benefit from our program, even though their test scores may not necessarily reflect it.

By construction, students in group (a) may overlap with the 7,272 students we had already identified, but students in group (c) can have no overlap with the existing sample. In total, this procedure enrolled 2,437 additional students.

This resulted in our final sample of 9,709 students.

B Supplementary tables and figures

This section contains supplementary tables and figures.

Table B.1 Balance test for main treatment sample. Referenced in Section 2.2.
Table B.2 Baseline academic covariates for RCT sample selected using each selection method. Referenced in Section 2.2.1.
Table B.3 Balance tests for tutor subtreatment. Referenced in Section 2.2.2.
Table B.4 Balance test for computer lab subtreatment. Referenced in Section 2.2.2.
Table B.5 Balance tests for incentive subtreatment. Referenced in Section 2.2.2.
Table B.6 Balance tests for selection treatment. Referenced in Section 2.2.1.

Table B.1: Balance table for main treatment.

	(1) Male	(2) Disadvantaged	(3) English-Medium	(4) Admin (z)	(5) Booklet (z)	(6) Quiz (z)	(7) Grit (z)
Treatment	0.00729	0.0150	0.0180	-0.0366	0.0483	-0.0169	-0.0945
	(0.0323)	(0.0226)	(0.0375)	(0.0492)	(0.0750)	(0.0590)	(0.0860)
Control Mean	0.3254	0.2249	0.4294	0.5223	0.5276	0.8598	0.6501
Control Sd	0.4688	0.4177	0.4953	0.7532	1.0272	0.9349	1.2024
p-value	0.8216	0.5072	0.6317	0.4576	0.5193	0.7750	0.2725
Step-down p-value (Romano-Wolf)	0.9750	0.9750	0.9750	0.9750	0.9750	0.9750	0.8861

Standard errors in parentheses, and clustered at school level. Table shows regressions of student characteristics on indicator for treatment. Statistics below table include control mean for the relevant variable, control standard deviation for the relevant variable, p-value on treatment, and step-down p-value for treatment applying the Romano-Wolf correction with 1,000 repetitions.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.2: Academic covariates for students selected by each selection method.

	(1) Admin Booklet Quiz	(2) Admin Quiz	(3) Booklet Quiz	(4) Admin Booklet	(5) Booklet	(6) Admin	(7) Quiz	(8) Grit
Exam scores	90.14	92.99	85.23	90.26	86.15	93.72	87.85	83.34
	(7.983)	(6.272)	(11.07)	(7.668)	(10.17)	(7.234)	(10.04)	(11.44)
Within-school z-score	1.654	1.835	1.504	1.750	1.565	1.864	1.586	1.526
	(0.580)	(0.560)	(0.717)	(0.620)	(0.777)	(0.614)	(0.696)	(0.773)
Admin (z)	0.561	0.810	0.174	0.584	0.244	0.887	0.379	0.0212
	(0.656)	(0.503)	(0.912)	(0.635)	(0.851)	(0.525)	(0.830)	(0.950)
Quiz Score (%)	63.11	63.85	65.37	58.33	63.43	54.92	66.93	63.13
	(16.83)	(16.79)	(19.39)	(19.33)	(18.77)	(18.30)	(13.35)	(20.24)
Booklet (z)	0.811	0.365	1.160	1.045	1.238	0.114	0.646	0.319
	(0.925)	(1.022)	(0.873)	(0.849)	(0.795)	(0.991)	(0.858)	(0.834)
Grit (z)	0.324	0.252	0.191	0.128	0.236	-0.0665	0.198	1.136
	(1.060)	(1.017)	(0.971)	(1.048)	(1.205)	(1.005)	(1.048)	(0.633)
Observations	495	556	269	224	183	318	200	245

Table shows mean (main) and standard deviation (parentheses) of various academic characteristics for students selected by each selection method.

Table B.3: Balance table for tutoring subtreatment.

	(1)	(2)	(3)	Diff. (1)–(2)	Diff. (1)–(3)	Diff. (2)–(3)
Male	0.318	0.364	0.297	-0.046*	0.020	0.066*
	(0.466)	(0.482)	(0.458)	(0.025)	(0.037)	(0.039)
Disadvantaged	0.237	0.254	0.205	-0.017	0.032	0.049
	(0.426)	(0.436)	(0.405)	(0.023)	(0.034)	(0.035)
English-Medium	0.419	0.467	0.503	-0.048*	-0.084**	-0.035
	(0.494)	(0.499)	(0.501)	(0.027)	(0.040)	(0.041)
Admin (z)	0.490	0.498	0.428	-0.008	0.062	0.070
	(0.780)	(0.747)	(0.902)	(0.041)	(0.064)	(0.065)
Booklet (z)	0.541	0.630	0.547	-0.089*	-0.006	0.083
	(0.981)	(1.010)	(1.022)	(0.053)	(0.079)	(0.083)
Quiz (z)	0.820	0.875	0.836	-0.055	-0.016	0.040
	(0.892)	(0.940)	(0.838)	(0.050)	(0.071)	(0.077)
Grit (z)	0.580	0.564	0.431	0.016	0.148	0.132
	(1.223)	(1.216)	(1.089)	(0.066)	(0.096)	(0.098)
Observations	800	610	195	1410	995	805

Table contains means for tutoring subtreatments and tests for significance in baseline characteristics. Column (1) contains means and standard deviations in parentheses for the no tutoring arm, Column (2) for the group tutoring arm, and Column (3) for the individual tutoring arm. The next 3 columns contain the difference between the columns described, with standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.4: Balance table for computer lab subtreatment.

	(1)	(2)	Diff. (1)–(2)
Male	0.336	0.332	0.004
	(0.473)	(0.471)	(0.026)
Disadvantaged	0.192	0.259	-0.067***
	(0.394)	(0.438)	(0.024)
English-Medium	0.458	0.443	0.014
	(0.499)	(0.497)	(0.027)
Admin (z)	0.530	0.468	0.063
	(0.768)	(0.788)	(0.043)
Booklet (z)	0.649	0.547	0.103*
	(1.005)	(0.993)	(0.055)
Quiz (z)	0.806	0.858	-0.051
	(0.836)	(0.930)	(0.050)
Grit (z)	0.520	0.570	-0.050
	(1.255)	(1.185)	(0.067)
Observations	459	1146	1605

Table contains means for computer lab subtreatments and tests for significance in baseline characteristics. Column (1) contains means and standard deviations in parentheses for the computer lab arm and Column (2) for the no computer lab arm. The final column contains the difference between the columns described, with standard errors in parentheses.
$^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

Table B.5: Balance table for incentive subtreatment.

	(1)	(2)	(3)	Diff. (1)–(2)	Diff. (1)–(3)	Diff. (2)–(3)
Male	0.353	0.329	0.316	0.024	0.037	0.013
	(0.478)	(0.470)	(0.465)	(0.030)	(0.028)	(0.029)
Disadvantaged	0.258	0.248	0.216	0.010	0.041	0.032
	(0.438)	(0.432)	(0.412)	(0.027)	(0.026)	(0.026)
English-Medium	0.461	0.417	0.462	0.044	-0.002	-0.046
	(0.499)	(0.493)	(0.499)	(0.031)	(0.030)	(0.030)
Admin (z)	0.493	0.490	0.475	0.002	0.018	0.016
	(0.771)	(0.824)	(0.758)	(0.050)	(0.046)	(0.048)
Booklet (z)	0.437	0.594	0.690	-0.156**	-0.253***	-0.096
	(0.978)	(0.978)	(1.018)	(0.061)	(0.060)	(0.061)
Quiz (z)	0.784	0.864	0.880	-0.080	-0.096*	-0.016
	(0.913)	(0.868)	(0.927)	(0.056)	(0.056)	(0.056)
Grit (z)	0.472	0.547	0.642	-0.075	-0.170**	-0.095
	(1.228)	(1.198)	(1.187)	(0.075)	(0.073)	(0.073)
Observations	532	504	569	1036	1101	1073

Table contains means for incentives subtreatments and tests for significance in baseline characteristics. Column (1) contains means and standard deviations in parentheses for the enforced incentives arm, Column (2) for the unenforced incentives arm, and Column (3) for the no incentives arm arm. The next 3 columns contain the difference between the columns described, with standard errors in parentheses.

$^{*}\ p < 0.10, ^{**}\ p < 0.05, ^{***}\ p < 0.01$.

Table B.6: Balance table for selection subtreatment.

Comparison	Male	Disadv-antaged	English-medium	Admin z	Booklet z	Quiz z	Grit z
1 vs. 2	−0.005	0.005	0.015	0.012	0.042	−0.034	−0.000
	(0.016)	(0.015)	(0.016)	(0.033)	(0.032)	(0.033)	(0.039)
1 vs. 3	−0.019	−0.018	0.110***	−0.090**	−0.061	−0.047	−0.093*
	(0.019)	(0.017)	(0.018)	(0.039)	(0.038)	(0.038)	(0.048)
1 vs. 4	−0.048**	−0.026	0.073***	−0.032	0.136***	0.020	−0.178***
	(0.020)	(0.019)	(0.020)	(0.042)	(0.042)	(0.042)	(0.049)
1 vs. 5	0.011	−0.015	−0.014	−0.053	0.052	−0.085**	−0.146***
	(0.020)	(0.019)	(0.020)	(0.042)	(0.041)	(0.041)	(0.049)
1 vs. 6	−0.055***	−0.023	0.085***	−0.053	0.029	−0.077**	−0.148***
	(0.018)	(0.017)	(0.018)	(0.038)	(0.038)	(0.038)	(0.046)
1 vs. 7	−0.026	−0.014	0.046**	−0.034	−0.075*	−0.089**	−0.079
	(0.020)	(0.018)	(0.020)	(0.041)	(0.041)	(0.041)	(0.050)
1 vs. 8	−0.042**	−0.042**	0.042**	0.022	0.002	0.004	−0.133***
	(0.018)	(0.017)	(0.018)	(0.037)	(0.037)	(0.037)	(0.045)
2 vs. 3	−0.013	−0.024	0.095***	−0.102***	−0.103***	−0.013	−0.093*
	(0.019)	(0.017)	(0.019)	(0.039)	(0.038)	(0.039)	(0.048)
2 vs. 4	−0.043**	−0.032*	0.058***	−0.044	0.094**	0.053	−0.178***
	(0.020)	(0.019)	(0.020)	(0.042)	(0.042)	(0.042)	(0.049)
2 vs. 5	0.016	−0.021	−0.029	−0.065	0.010	−0.051	−0.146***
	(0.020)	(0.019)	(0.020)	(0.042)	(0.041)	(0.041)	(0.049)
2 vs. 6	−0.050***	−0.028*	0.070***	−0.065*	−0.013	−0.043	−0.148***
	(0.018)	(0.017)	(0.018)	(0.038)	(0.038)	(0.038)	(0.046)
2 vs. 7	−0.021	−0.020	0.031	−0.046	−0.117***	−0.056	−0.079
	(0.020)	(0.018)	(0.020)	(0.041)	(0.041)	(0.041)	(0.050)
2 vs. 8	−0.037**	−0.047***	0.026	0.010	−0.040	0.037	−0.133***
	(0.018)	(0.017)	(0.018)	(0.037)	(0.037)	(0.037)	(0.045)
3 vs. 4	−0.030	−0.008	−0.037	0.058	0.197***	0.066	−0.085
	(0.022)	(0.021)	(0.022)	(0.047)	(0.046)	(0.047)	(0.057)
3 vs. 5	0.030	0.003	−0.124***	0.037	0.113**	−0.038	−0.053
	(0.022)	(0.021)	(0.022)	(0.046)	(0.046)	(0.046)	(0.057)
3 vs. 6	−0.037*	−0.005	−0.025	0.037	0.090**	−0.030	−0.055
	(0.021)	(0.019)	(0.021)	(0.044)	(0.043)	(0.043)	(0.054)
3 vs. 7	−0.008	0.004	−0.064***	0.056	−0.014	−0.043	0.014
	(0.022)	(0.020)	(0.022)	(0.046)	(0.045)	(0.046)	(0.058)
3 vs. 8	−0.024	−0.023	−0.068***	0.112***	0.063	0.050	−0.040
	(0.020)	(0.019)	(0.020)	(0.043)	(0.042)	(0.042)	(0.054)
4 vs. 5	0.059**	0.011	−0.087***	−0.021	−0.084*	−0.104**	0.032
	(0.024)	(0.022)	(0.023)	(0.049)	(0.049)	(0.049)	(0.058)
4 vs. 6	−0.007	0.003	0.012	−0.021	−0.107**	−0.097**	0.030

Comparison	Male	Disadv-antaged	English-medium	Admin z	Booklet z	Quiz z	Grit z
	(0.022)	(0.021)	(0.022)	(0.046)	(0.046)	(0.046)	(0.055)
4 vs. 7	0.022	0.012	−0.027	−0.002	−0.211***	−0.109**	0.099*
	(0.023)	(0.022)	(0.023)	(0.049)	(0.048)	(0.048)	(0.058)
4 vs. 8	0.006	−0.015	−0.032	0.053	−0.134***	−0.016	0.045
	(0.022)	(0.020)	(0.022)	(0.046)	(0.045)	(0.045)	(0.054)
5 vs. 6	−0.066***	−0.008	0.099***	0.000	−0.023	0.008	−0.002
	(0.022)	(0.020)	(0.022)	(0.046)	(0.046)	(0.046)	(0.055)
5 vs. 7	−0.037	0.001	0.060***	0.019	−0.127***	−0.005	0.067
	(0.023)	(0.021)	(0.023)	(0.048)	(0.048)	(0.048)	(0.058)
5 vs. 8	−0.053**	−0.026	0.055**	0.075*	−0.050	0.088**	0.013
	(0.022)	(0.020)	(0.022)	(0.045)	(0.045)	(0.045)	(0.054)
6 vs. 7	0.029	0.009	−0.039*	0.019	−0.104**	−0.012	0.069
	(0.022)	(0.020)	(0.022)	(0.045)	(0.045)	(0.045)	(0.056)
6 vs. 8	0.013	−0.018	−0.044**	0.074*	−0.027	0.081*	0.015
	(0.020)	(0.019)	(0.020)	(0.042)	(0.042)	(0.042)	(0.052)
7 vs. 8	−0.016	−0.027	−0.005	0.056	0.077*	0.093**	−0.054
	(0.021)	(0.020)	(0.021)	(0.045)	(0.044)	(0.044)	(0.055)

Table presents comparisons of means of demographic variables for the set of students who were considered by each selection method. These means are calculated for *all* 9,709 screened students, not just those selected; as such, they should be balanced across covariates. Table contains both differences in means and standard errors in parentheses. Here, (1) refers to *Admin+Booklet+Quiz*, (2) to *Admin+Quiz*, (3) to *Booklet+Quiz*, (4) to *Admin+Booklet*, (5) to *Booklet*, (6) to *Admin*, (7) to *Quiz*, and (8) to *Grit*.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.