NETWORK LISER

Discussion Paper Series

IZA DP No. 18381

February 2026

Developing Math Talent Worldwide: Evidence from a Global RCT

Ruchir Agarwal
Harvard Kennedy School

Patrick Gaule
University of Bristol and IZA@LISER

IZA NETWORK @ LISER
ISSN: 2365-9793

IZA DP No. 18381 February 2026

Developing Math Talent Worldwide: Evidence from a Global RCT*

Abstract

Exceptional talent accounts for a disproportionate share of innovation, yet many individuals with exceptional ability may never realize their potential. Whether expanding access to advanced training generates learning gains remains an open question. We study this using a randomized controlled trial with 620 highly gifted students from 44 countries, nominated by national Olympiad organizations. Participants were randomly assigned either to an 18-week advanced combinatorics course by Art of Problem Solving or to independent study using equivalent materials. Assignment to the course increased final-exam performance by 0.16 standard deviations. Engagement varied widely: roughly half of assigned students participated minimally, and baseline characteristics explain little of this variation ($R^2 \approx 0.10$). Using random assignment as an instrument for engagement, we estimate learning gains of 0.66 standard deviations among fully engaged students. Among those who later competed in the International Mathematical Olympiad, students assigned to the course performed better on combinatorics problems. Overall, access to advanced training yields large gains when engagement is sustained, but access alone does not reliably induce engagement.

JEL classification

I21, J24 ,O31

Keywords

exceptional talent, gifted education, randomized controlled trial, student engagement, human capital, mathematics education, olympiad training

Corresponding author

Patrick Gaule

patrick.gaule@bristol.ac.uk

* We appreciate helpful comments from Deivis Angeli, Toman Barsbai, Simon Coyle, Glenn Ellison, Santiago Saavedra, and Josh Taylor. We are indebted to Zurab Abramishvili, Vladimir Bragin, and Yuyan Jiang for their help in executing the field experiment. João Francisco Gomes Marques provided excellent research assistance. We are grateful to Brent Nicklas, Rebecca Sodervick, Deven Ware, and Richard Rusczyk from Art of Problem Solving (AoPS), and Eljakim Schrijvers at Eljakim Information Technology Limited, for their help in implementing aspects of the program. We appreciate the cooperation of many partners in national Olympiad organizations, with special mention of Sarfraz Ahmad, Richard Eden, Claudio Landim, Syamil Ahmad Shakir, and Otgonbayar Uuye. The field experiment was reviewed and approved by the School of Economics Research Ethics Committee at the University of Bristol, and pre-registered in the AEA RCT registry (AEARCTR-0011178). The program was implemented by Global Talent Fund, a US nonprofit we co-founded. AoPS provided in-kind support related to course delivery, including discounted pricing, scheduling accommodations, and access to administrative data, but had no role in the research design or the interpretation of results. We acknowledge financial support from Carina Initiatives, Polynera and Shaffer, Founders Pledge, and Schmidt Futures.

1. Introduction

Exceptional talent drives a disproportionate share of innovation and scientific discovery, making the study of learning at the top of the ability distribution central to long-run economic growth. Yet we have little causal evidence on the conditions under which such learning occurs, or on the constraints that limit its development. Most education systems are designed to serve the median student, while the processes governing learning among individuals with exceptional ability—whose learning dynamics, opportunities, and constraints may differ sharply—remain poorly understood.

One central concern for innovation and growth is that many individuals with exceptional ability may never realize their potential. Advanced problem-solving skills typically develop through exposure to challenging material, expert instruction, and sustained practice—inputs that are often geographically concentrated in a small number of elite schools and training centers. As a result, even students with exceptional ability and strong motivation may fail to reach the frontier if they live far from high-quality training environments.[1] This creates the possibility of large, socially costly losses in unrealized talent, not because ability is scarce, but because access to the conditions that foster advanced learning is unevenly distributed.

The career of Maryam Mirzakhani illustrates both the promise and fragility of frontier learning. As a middle-school student in Tehran in the early 1990s, Mirzakhani did not initially excel in mathematics and was more interested in literature and debate. Her subsequent development coincided with Iran's expansion of programs for mathematically gifted students, including national Olympiad training camps and university workshops. Within a few years, she excelled at the International Mathematical Olympiad (IMO) and later earned the Fields Medal—the highest honor in the field. Her experience suggests how rapidly learning can accelerate when access to advanced training is combined with a supportive environment. At the same time, it raises a broader question: *if access to advanced training were expanded to students with exceptional ability at scale, how much learning would occur, and under what conditions would large gains be sustained?*[2]

[1]Birth location remains a powerful determinant of whether exceptional ability develops. For instance, Africa—nearly 20 percent of global population—has won only three International Mathematics Olympiad (IMO) gold medals and zero Physics Nobel prizes. These patterns suggest that access to training, not just underlying ability, shapes outcomes at the frontier.

[2]Maryam Mirzakhani's early education and rise through Iran's emerging Olympiad system are described in Beheshti (2018). As a middle-school student at Farzanegan School for gifted girls in Tehran, Maryam initially did not excel in mathematics. After scoring poorly on a test, she reportedly tore up her exam paper in frustration and declared she would never study math again. Her turning point came when Iran launched new programs for mathematically gifted students, including national Olympiad training camps and a summer workshop at Sharif University. Alongside her close friend Roya Beheshti, she became one of the first two female students on Iran's national Math Olympiad team, earning gold medals in 1994 and 1995, the latter with a perfect score. She later completed her Ph.D. at Harvard and went on to become the first woman to win the Fields Medal in 2014.

Around the world, many students with exceptional ability face similar constraints.[3] Progress in competition-level mathematics requires mastering problem-solving techniques that are rarely taught in standard curricula and typically demand sustained interaction with highly specialized instructors and similarly prepared peers. As a result, even students who have demonstrated exceptional aptitude and motivation may struggle to continue progressing once they exhaust locally available instruction.

Online education offers a potential solution to these constraints. Traditional education systems face structural limits in serving students at the upper tail of the ability distribution: few schools have teachers equipped to teach competition-level mathematics, and even where such teachers exist, the scarcity of similarly prepared peers often prevents the formation of viable classes. By connecting students to expert instructors and advanced curricula regardless of geography, online instruction could, in principle, make access to advanced training far more scalable than traditional in-person models.

Yet how much access to high-quality online instruction alone can raise learning for high-ability students—and under what conditions—remains an open question. Most existing research on online education focuses on mainstream student populations. Much less is known about whether scalable online instruction can effectively serve students with exceptional ability (whose learning needs, motivations, and persistence patterns may differ sharply), or what determines whether such students persist.[4,5]

This paper addresses that gap. We conduct the first randomized controlled trial of high-quality online education for top mathematics students, spanning 620 participants across 44 countries. Participants were nominated by national Olympiad organizations and drawn from among the world's most mathematically able students.[6] The intervention provided free access to an 18-week advanced combinatorics course from Art of Problem Solving (AoPS), a leading mathematics training platform whose alumni include most members of the U.S. International Mathematical Olympiad team. Students in the control group received an equivalent AoPS

[3]Consider Kevin (a pseudonym), a 16-year-old from Mindanao in the southern Philippines who participated in our study. Like many high-ability students outside major urban centers, he has competed in local mathematics contests since age ten and aspires to earn admission to a top university or international scholarship. To participate in national Olympiad training, he would need to travel to Manila—a journey of two to three days by ferry.

[4]Throughout the paper, we use the term persistence to refer to a student's ability to sustain effort over time when working on demanding, advanced problem-solving tasks. Empirically, we measure persistence through engagement in the course, defined as the number of weeks or months in which students actively participated by attempting and submitting problem sets.

[5]Students who compete in math Olympiads—with demonstrated ability, prior investment, and strong academic incentives—represent a likely case for online education to succeed. If persistence constrains learning even here, it might bind more severely in less selected populations.

[6]Participation in Olympiad competitions itself reflects a dual self-selection: students must be both passionate about mathematics and confident enough in their abilities to voluntarily tackle advanced problems well beyond standard curricula. Among this group of 10,000 or more national participants (numbers vary by country), our participants typically ranked in the top 20, placing them in approximately the top 0.2% of students who have self-selected into competitive mathematics.

e-book and structured study plan.

To measure and incentivize learning, we administered a dedicated exam approximately 20 weeks after the program began. Some students in both the treatment and control groups did not sit the final exam. Rather than treating this nonparticipation as missing data, we interpret it as informative about learning and engagement, and define outcomes accordingly.[7] Cash prizes of USD 500 were awarded to the top third of participants, with incentives announced in advance to encourage sustained effort. Students also received certificates and access to a global learning community, ensuring that the experimental design captured meaningful engagement even among a highly selected and motivated population.

Using data from this randomized controlled trial, we estimate the causal effect of high-quality online instruction for top mathematics students. To address endogenous participation, we employ a two-stage least squares framework that instruments course participation with random assignment.

We report four main findings:

1. **Access to advanced training raises learning at the frontier.** Intention-to-treat estimates show that assignment to advanced training increased final-exam performance by 0.16 standard deviations. These average effects are sizable for a scalable intervention, but mask substantial heterogeneity in engagement and learning outcomes.

2. **But engagement varies widely and is difficult to predict.** Roughly half of students assigned to the course participated minimally, while only 15 percent remained fully engaged throughout the 18-week course. Observable baseline characteristics explain less than 10 percent of the variation in engagement ($R^2 \approx 0.10$). As a result, engagement plays a central role in shaping overall learning impacts.

3. **For students who persist, learning gains are large and extend to external assessments.** Each additional month of active participation increased final-exam performance by 0.14–0.16 standard deviations, implying learning gains of approximately 0.66 standard deviations for students who remained fully engaged. These gains are large by educational standards, particularly given that participants were already among the top mathematics students in their countries. Consistent with these effects, among a subset of 39 students who later competed in the International Mathematical Olympiad, assignment to the course substantially improved performance on the combinatorics problems, with no effects on participation or on non-combinatorics topics. Despite the limited sample size, this provides rare experimental evidence that access to advanced training can improve performance on frontier external assessments.

[7]Section 5.5 discusses robustness to alternative treatments of exam nonparticipation.

4. **Expanding access to advanced training does not mechanically equalize opportunity.** Learning gains are larger for students with higher initial ability and for those in capital cities and countries with stronger traditions of advanced mathematics training. We find no evidence of relatively stronger effects for participants in more peripheral locations or weaker Olympiad environments. These patterns suggest that expanded access to advanced training may complement existing systems of preparation and support rather than substitute for them.

Taken together, these findings point to a central tension. When access to advanced training is no longer the barrier, persistence becomes the binding constraint—even among the world's most talented and motivated students. Online education can generate large learning gains for those who remain engaged, but many do not, and predicting who will persist remains difficult. As a result, the promise of scalable access to advanced training is only partially realized in practice. Online education thus appears most effective when combined with conditions that support persistence, rather than as a standalone solution for scaling elite talent development. As AI-driven tools further lower the marginal cost of delivering advanced instruction, our results suggest that the binding constraint in scaling elite talent may increasingly lie in persistence rather than in access itself.

This paper contributes to several strands of literature. First, the study contributes to the emerging literature on the role of talent in innovation and economic growth (e.g., Hsieh et al. 2019; Bell et al. 2019). Agarwal and Gaule (2020) document the phenomenon of "invisible geniuses"—talented individuals whose potential remains unrealized due to lack of access to appropriate educational opportunities—and estimate that alleviating this constraint could substantially accelerate knowledge production. Related work shows that performance in mathematics Olympiads strongly predicts subsequent achievement in science and innovation: Olympiad medalists account for roughly 10 percent of the world's most prestigious science prizes, at more than twenty times the rate of top university graduates (Agarwal, Gaule, & Jiang 2025). However, this literature focuses on outcomes conditional on having already reached Olympiad-level performance, leaving open the question of how individuals acquire the advanced problem-solving skills required to reach that frontier.

Second, it adds to the growing body of experimental evidence on technology-aided instruction and online learning. Prior randomized evaluations document learning gains from computer-assisted instruction in primary and secondary schools (Banerjee et al. 2007; Muralidharan et al. 2019) and from online or blended courses in higher education (Figlio et al. 2013; Bettinger et al. 2017; Cacault et al. 2021; Lichand et al. 2022; Kofoed et al. 2024). This literature establishes that digital delivery can substitute for, or complement, traditional instruction in mainstream educational settings. However, it provides little evidence on whether scalable online instruction can effectively develop advanced skills at the extreme upper tail of the ability distribution. We provide the first experimental evidence on online education for students at this upper tail—a group that plays a disproportionate role in innovation but has received little empirical attention.

A recent study by Adajar, Duflo, Ellison, Ellison, and Kannan (2025) also collaborates with Art of Problem Solving (AoPS), but examines a very different segment of the education landscape. Their randomized evaluation studies how relatively strong seventh-grade students in government schools in the Indian state of Tamil Nadu engage with a self-paced AoPS pre-algebra course, and how design features such as tutoring support or access to computer labs affect engagement in a low-resource context. While these students are high performers within their schools, the authors note that they typically begin with mathematics knowledge comparable to the median American seventh grader.

Third, the study contributes to the literature on interventions for gifted and high-ability learners. Prior research shows that tracking and targeted programs can improve outcomes for high-achieving students (Duflo et al. 2011; Card and Giuliano 2014, 2016), but these interventions typically rely on in-person instruction, specialized institutions, and small peer groups, limiting their scalability. As a result, this literature offers limited guidance on whether elite talent can be developed at scale, or on which constraints bind when instructional access is expanded. Our results show that high-quality virtual instruction can generate large learning gains for gifted students who remain engaged, while also highlighting important limits to its reach when engagement is not sustained.

Fourth, the paper contributes to understanding the mechanisms through which online education affects learning. A large literature documents substantial drop-off in online course participation, with completion rates in massive open online courses often below 10 percent (Gong et al. 2021). Most studies treat completion or persistence as outcomes in their own right. By contrast, our instrumental variables approach separates the effect of access to instruction from the effect of sustained engagement. We show that average treatment effects mask much larger returns to engagement, and that persistence remains both limited and difficult to predict—even in a highly selected and strongly incentivized sample, precisely the setting in which sustained engagement might have been expected ex ante.

The present study addresses this gap by experimentally evaluating a key upstream input in the talent development pipeline: access to specialized training in advanced problem-solving. We show that high-quality virtual instruction can generate large learning gains for students who remain engaged—including on external competitive assessments—while also highlighting substantial limits to its reach when engagement is not sustained. In this way, our findings speak to both the potential and the practical constraints of using technology to expand access to advanced training in the pre-Olympiad phase.

The remainder of the paper proceeds as follows. Section 2 provides background on mathematics competitions, competition training, and the Art of Problem Solving. Section 3 describes the experimental design and program implementation, including recruitment procedures, the intervention, and outcome measurement. Section 4 presents the data sources, baseline characteristics, and balance tests. Section 5 reports the main results, including intention-to-treat effects, patterns of engagement and persistence, heterogeneous treatment effects, and instrumental variable estimates. Section 6 discusses the findings and their implications for the delivery and support of advanced mathematical instruction.

2. Background

This section describes the international mathematics competition system and the training environments through which advanced problem-solving skills are typically developed.

2.1. The International Mathematical Olympiad

The International Mathematical Olympiad (IMO) is the most prominent international mathematics competition for secondary school students. Held annually since 1959, it brings together teams of up to six students from more than 100 countries. Although participants represent national teams, all examination work is completed and scored individually. The competition consists of two examination days, each featuring three newly designed problems drawn from algebra, number theory, combinatorics, and geometry, with individual problems scored on a 0–7 scale. Gold, silver, and bronze medals are awarded based on individual total scores. Because the problem sets are newly created and centrally administered each year, IMO performance provides a standardized measure of advanced problem-solving ability among a highly selected group of students. Success at this level requires not only exceptional ability but sustained preparation over several years.

2.2. The Competition System

Mathematics competitions for secondary school students are held in many countries worldwide, with a history dating back to the late nineteenth-century Austro-Hungarian Empire (de Losada and Taylor 2022). These competitions aim to promote mathematical learning, develop problem-solving skills, and identify exceptional talent. They typically follow a pyramidal structure: large numbers of students participate at local or regional levels, progressively smaller groups advance to national competitions, and the highest performers proceed to international events such as the International Mathematical Olympiad. Beyond selection, competition systems serve as a primary mechanism through which mathematical ability is identified and developed.

2.3. Students and Their Motivations

Mathematically gifted secondary school students often seek opportunities to learn mathematics beyond the standard school curriculum. Some pursue advanced topics typically taught at the undergraduate level.[8] Others go deeper into "elementary" mathematics—combinatorics, number theory, algebra, and geometry—which reveal substantial depth when approached through competition-style problems.

Students participate in competitions for a variety of reasons: the intellectual challenge of solving difficult problems, opportunities to meet like-minded peers, international travel, and recognition that may help in university admissions. In China, India, and Brazil, for instance, outstanding performance in national mathematics competitions can lead to direct admission to prestigious universities. Some American and British universities also consider competition results in their admissions decisions, including for international applicants.

2.4. Training for Mathematics Competitions

Success in mathematics competitions typically requires extensive preparation and sustained practice. In addition to general mathematical aptitude, participants must master a broad range of problem-solving techniques and learn to apply them flexibly to unfamiliar problems under time constraints. Because standard school curricula rarely cover this material, most students depend on extracurricular training—and access to such training varies widely across countries.

In a few places, students attend specialized secondary schools that offer intensive preparation for high-level competitions.[9] Much more often, students train outside the formal school system. One common approach is the mathematics circle, in which an instructor—often a university professor—meets regularly with students to work through challenging problems and solution strategies. Other students prepare on their own, work with private tutors, or rely

[8]The Advanced Placement (AP) program of the U.S. College Board offers university-level curricula and examinations to high school students in mathematics and other subjects. Students who perform well may receive college credit, reducing the length and cost of college. Some universities also consider AP results in admissions or financial aid decisions.

[9]Some countries have specialized high schools catering to talented students, including boarding facilities where needed. Examples include the Mathematical Grammar School in Belgrade, Serbia; the Komarovi Physics and Mathematics School in Tbilisi, Georgia; and the International Computer Science High School of Bucharest in Romania. Outside former communist countries, such specialized instruction is either unavailable or accessible only through expensive private institutions.

on online materials, while many countries also organize intensive training camps for their top performers. Overall, access to high-quality preparation remains unevenly distributed, and even when access is expanded, sustained engagement often depends on the surrounding training environment and available support.

2.5. Art of Problem Solving

Art of Problem Solving (AoPS) is a U.S.-based educational company specializing in mathematics education for gifted and highly motivated students. Originally a textbook publisher, AoPS has expanded to offer online and in-person instructional programs. Its virtual course platform, AoPS Online, provides live, interactive classes taught by experienced instructors, complemented by challenging homework and written problems completed asynchronously.

AoPS online courses are widely used in mathematics competition training in the United States and have a long track record of delivering standardized, high-level instruction to advanced students, including most members of the US IMO teams in recent years. Outside the United States, however, participation in AoPS online courses has been more limited. This may reflect the relatively high cost of enrollment and scheduling constraints, as most classes are offered in U.S. evening time slots. This limited international reach creates a natural setting to study whether expanding access to high-quality online instruction can accelerate learning among high-ability students who would otherwise lack such opportunities.

3. Field Experiment and Implementation

3.1. Recruiting Participants

Our study focuses on the effects of distance learning among students with exceptional mathematical ability, as identified through national mathematics competitions. Such students are rare, and the course evaluated in the experiment presupposes substantial prior exposure to competition-style mathematics and covers material well beyond standard secondary-school curricula. As a result, the intervention is appropriate only for a narrow group of learners at the top of the ability distribution, making recruitment a practical challenge. At the same time, these students represent a population for whom access to advanced training is often limited by geography, making them a natural group in which to study whether expanding access can accelerate learning.

To identify students for whom the course content was well matched, we collaborated with national mathematics olympiad organizations in 44 countries, with a focus on developing countries, requesting that they nominate high-performing students based on results in national competitions.[10]

At the recruitment stage, students were informed that participation in the program offered several potential benefits. These included access to high-quality training materials from Art of Problem Solving, the possibility of enrolling in an 18-week AoPS online course on counting and probability, and formal recognition through certificates of participation and excellence. In addition, participants were informed that cash prizes of up to $500 would be awarded to approximately one third of students, conditional on completing the final assessment and based on relative performance. These features were intended to encourage participation and engagement throughout the study period.

Interested students who provided informed consent and completed a baseline survey were asked to take a selection test prior to randomization. The selection test served two purposes: first, as an eligibility screen to verify that participants had the prerequisite level of mathematical preparation; and second, to provide a standardized pre-intervention measure of mathematical skills and knowledge used in the empirical analysis.

3.2. Intervention

Half of the study participants were randomly assigned to receive access to an advanced mathematics course.[11] The course was delivered online by Art of Problem Solving (AoPS), a specialized provider of instructional materials and virtual courses for high-ability students. The course offered in the experiment was identical to AoPS's commercially available offerings and drew on the provider's established approach to teaching advanced mathematics in an online setting. Course delivery, grading, and student engagement tracking followed AoPS's standard operating procedures and were implemented uniformly across cohorts.

Instruction combined weekly live interactive sessions with graded homework assignments. The course ran for 18 weeks, with one 90-minute session per week delivered through a text-based online interface. Sessions emphasized problem solving, with instructors presenting sequences of increasingly challenging problems and providing explanations after students had attempted solutions. Teaching assistants provided support as needed. Outside of live sessions, participants completed weekly homework consisting of both automatically graded challenge problems and written solutions that received individualized feedback. A course-specific online forum supported asynchronous discussion. A fully engaged student typically devoted several hours per week to sessions and assignments over the 18-week program.

[10] Recruitment varied across countries depending on country size, institutional arrangements, and local take-up rates; Appendix B provides further details.

[11] The randomization procedure is described in Section 3.3.

The course focused on combinatorics, the branch of mathematics concerned with counting. We selected combinatorics for several reasons: it is one of the four core areas assessed in mathematics competitions at both national and international levels; it is rarely taught in depth within standard secondary-school curricula; it has broad applications in computer science, statistics, and related fields; and it aligns well with the AoPS instructional approach, which emphasizes learning through problem solving. The course assumed prior mastery of algebra and familiarity with introductory counting and probability concepts.[12]

Participants assigned to the control group were provided free access to an Art of Problem Solving e-book and an accompanying study plan. The e-book, which also served as the textbook for course participants,[13] covered the same subject matter as the course. The study plan outlined suggested readings and exercises drawn from the book but did not include live instruction, graded assignments, or feedback. Appendix B documents that independent enrollment in similar AoPS courses during the study period was extremely rare among study participants.

In summary, the intervention created a contrast between access to a structured, instructor-led online course in advanced combinatorics and access to self-directed study materials covering the same content. Appendix B provides additional detail on course content and instructional format.

3.3. Study Design and Randomization

Upon completing a consent form, passing the selection test, and finishing the baseline survey, students were enrolled in the study and randomly assigned to treatment and control groups. Students were grouped into strata (matched quadruplets) based on their selection test score.[14] Within each stratum, two students were randomly assigned to receive access to the virtual course and two to the control condition.

This stratified randomization approach was chosen to maximize statistical power by balancing treatment assignment on a strong predictor of learning outcomes (Bruhn and McKenzie 2009; Duflo, Glennerster, and Kremer 2007). In addition, because the analysis includes stratum fixed effects, using strata of four rather than matched pairs reduces the number of observations lost to attrition at endline. Following recent guidance emphasizing the advantages of larger strata in the presence of attrition, we therefore implemented randomization within matched

[12] Topics covered in the course included inclusion–exclusion, recursion, conditional probability, generating functions, and introductory graph theory, as well as standard combinatorial tools such as permutations, combinations, Pascal's triangle, combinatorial identities, and the Binomial Theorem.

[13] Course participants received automatic access to the e-book as part of their enrollment.

[14] When more than four students had identical test scores, we used country IMO rankings as an additional stratification variable.

quadruplets (Bai et al. 2024; McKenzie 2022).

3.4. Measurement of Learning Outcomes

We measured learning outcomes using a dedicated final exam administered approximately 20 weeks after the program began. The exam was designed to assess problem-solving ability in advanced combinatorics and related topics covered in the course and serves as the primary learning outcome in the analysis.

To ensure appropriate difficulty and relevance for the study population, we commissioned an expert mathematician to design a set of original problems specifically for this assessment. The exam was administered uniformly to students in both treatment and control groups and was independent of AoPS course materials. Appendix B describes the design, administration, grading, and integrity safeguards of the final exam in detail.

3.5. Implementation

The randomized trial was implemented by Global Talent Fund (formerly called Global Talent Network), a nonprofit organization we cofounded. The trial was registered in the AEA RCT Registry in April 2023.

Students participated in one of four cohorts. The first cohort began in April 2023, followed by cohorts starting in September 2023, March 2024, and September 2024. In each case, students participated in the program for approximately five months, culminating in the final assessment.

The first cohort, comprising 77 students from three countries (the Philippines, Malaysia, and Mongolia), served as a pilot to inform the implementation of subsequent cohorts.[15] From the second cohort onward, we implemented the selection test as a screening device and baseline measure of ability. In addition, we invited prospective participants to take part in a two-day AoPS workshop prior to enrollment.[16] Figure 1 shows the geographic distribution of participants across cohorts.

AoPS implemented the instructional component of the intervention using its standard course infrastructure. AoPS had no role in the study design, randomization, outcome construction, or statistical analysis. Learning outcomes were assessed using exams designed and graded independently of AoPS, and all empirical analysis was conducted by the authors.

[15] The pilot cohort featured simplified recruitment procedures and larger cash incentives, with prizes of $1,000 awarded to top performers.

[16] The workshop provided students with a sample of the instructional approach used in the course and allowed us to observe engagement in a low-stakes setting, while ensuring that all participants received a tangible benefit from participation.

4. Data

This section describes the data we use for the experimental analysis. We draw on several sources: administrative records from the registration process and selection tests, a baseline survey of participants, detailed AoPS course engagement data, and results from the final exams that we administered. After describing these data sources, we present descriptive statistics on participant characteristics and test for balance across the treatment and control groups.

4.1. Sources

Registration and Baseline Survey. As part of the study registration process, we collected basic participant information, including name, contact details, and country of residence. Participants also completed a baseline survey covering demographic characteristics, socioeconomic background, prior competition experience, and English proficiency. Following registration, participants were invited to take part in a selection test.

Selection Test. Beginning with the second cohort, we administered a standardized selection test prior to randomization to measure baseline mathematical ability. For the first cohort, which served as a pilot, baseline ability was proxied using available national competition outcomes or training camp results, standardized within country. For a small number of participants for whom no such measure was available, we assigned the sample median. Details are provided in Appendix B.

Course Engagement Data. As part of its standard practice, Art of Problem Solving (AoPS) systematically records participant activity using platform-generated administrative data, allowing us to observe engagement on a weekly basis. Engagement data primarily reflects completion of graded problem sets rather than attendance at live sessions, is available on a weekly basis, and is color-coded as follows: blue (mastery), green (pass), orange (engaged but struggled), and red (non-engagement). Through our collaboration with AoPS, we obtained teacher-level access to participation metrics for the study participants. In this paper, we focus on the number of weeks during which students remained engaged in the course. Because our interest lies in measuring sustained effort rather than mastery, students coded as 'engaged but struggled' are treated as engaged.

Book Engagement Data. Students assigned to the control group received access to an AoPS ebook covering the same mathematical content as the virtual course. As an interactive online document rather than a static PDF, the ebook platform automatically tracks which content fragments each student viewed. Through our collaboration with AoPS, we obtained fragment-level engagement data for all control-group participants. The ebook contains 134 fragments, and we measure engagement as the share of fragments viewed. This metric captures breadth of exposure to the material, though it does not distinguish between superficial clicking and careful reading.

Final Exam. Final exam results are available for all participants who sat the exam. As students participated in one of four cohorts, there are four distinct final exams. To account for differences in test difficulty across cohorts, we standardize scores to have mean zero and variance one. Students who did not take the final exam (around 40% of the sample) are coded as having an outcome of zero in the main analysis, treating exam non-participation as an informative outcome rather than as missing data. Attrition rates are similar across treatment and control groups. We examine the determinants of non-participation and assess robustness to alternative treatments of missing data in Section 5.5.

International Mathematical Olympiad Outcomes. As a secondary outcome, we link study participants to publicly available records from the International Mathematical Olympiad (IMO), which report participation and problem-level scores for each competitor. These data provide an external, independently administered measure of advanced mathematical problem-solving ability. Because IMO participation and timing vary across countries and cohorts, we examine outcomes in the first IMO occurring after the intervention and account for timing differences using cohort fixed effects. Details on data matching and problem classification are provided in Appendix B.10.

4.2. Baseline Characteristics and Balance Tests

Table 1 presents summary statistics on the baseline characteristics of study participants for the pooled sample and separately for the control and treatment groups.

Baseline Characteristics. Study participants were in their mid-teens, with an average age of 16.5 years, and 23% were female. They came from relatively advantaged socio-economic backgrounds: 76% had at least one parent with a university degree, and more than half had traveled internationally outside of competitions. Most participants were well-positioned to learn remotely, with 95% having internet access at home and 90% having access to a computer. Moreover, more than 90% reported intermediate or advanced proficiency in English.

Participants had substantial prior exposure to mathematics competitions, with an average of 4.7 years of experience, and performed well on the selection test, solving on average five of the nine problems presented. However, there was considerable variation in participants' learning environments. Roughly half were based in their country's capital city, while the others lived elsewhere. Participants came from countries with varying competition experience, ranging from long-established programs (e.g., Bulgaria, Mongolia) to beginners (e.g., Ethiopia).

Balance Tests. The final column of Table 1 reports p-values from tests of mean differences between the treatment and control groups for each variable. Baseline characteristics were generally balanced between the treatment and control groups, as expected under random assignment. The treatment group was slightly older and more likely to have traveled internationally. However, variables more likely to be predictive of final exam outcomes—such as performance on the selection test and years of competition experience—were closely balanced.

5. Results

5.1. Intention-To-Treat Effects

Specification. We estimate the intention-to-treat (ITT) effect of the intervention using the following specification:

$$Y_i = \beta\, Treat_i + \alpha_i + \gamma^\top \mathbf{X}_i + \varepsilon_i, \tag{1}$$

where Y_i denotes the final exam score, normalized to have mean zero and variance one; $Treat_i$ is an indicator for assignment to the treatment group; α_i captures fixed effects for small strata; and $\mathbf{X}_i$ is a vector of additional baseline controls. Our preregistered specification includes stratum fixed effects α_i[17] but excludes additional controls. As a robustness exercise, we present these preregistered results alongside specifications that include baseline controls $\mathbf{X}_i$ as well as a parsimonious regression of outcomes on the treatment indicator $Treat_i$ only.

Results. Table 2 reports the intention-to-treat effects on learning. Students assigned to the virtual course score 0.165 standard deviations higher on the final exam (column 2), a statistically significant effect at the 5 percent level. Estimates are very similar in the specification without strata fixed effects (0.150 standard deviations), and the specification including baseline controls yields a slightly larger effect of 0.173 standard deviations.

5.2. Heterogeneous Effects

We examine whether treatment effects vary across student subgroups. Table 3 reports estimates separately by training environment—specifically, whether the student's country has a strong national mathematical olympiad tradition and whether they are located in the capital—and by baseline ability, proxied by the selection test. These analyses provide evidence on potential mechanisms underlying the intervention and inform the design and targeting of related programs.

By the Quality of the Environment. When we designed the experiment, we hypothesized

[17]The stratum fixed effects correspond to the matched quadruplets used in randomization

that the intervention might be particularly beneficial for students with more limited training opportunities, potentially leveling the playing field. We operationalize training environment using each country's recent performance in the International Mathematical Olympiad (IMO), measured by its most recent IMO rank. Students from countries with stronger IMO performance exhibit an estimated positive effect of 0.196 standard deviations, compared to 0.093 standard deviations for students from countries with weaker IMO performance. However, this difference is not statistically significant (p = 0.521), providing no evidence that the treatment effect varies systematically with the strength of a country's training environment. Similarly, the estimated effect is 0.269 standard deviations for students in capital cities and 0.048 standard deviations for those outside the capital, with the difference again not statistically significant (p = 0.172). Overall, the data do not provide evidence of systematically larger effects for students in weaker environments. If anything, the point estimates are larger for students with stronger existing training environments, although these differences are not statistically significant.

By Ability. We also examine heterogeneity by prior ability. While we recruited high achievers with solid foundations, there remains substantial variation within the sample, as reflected in selection-test scores. The estimated effect is 0.254 standard deviations for students scoring above the median on the selection test and 0.006 standard deviations for those scoring below the median. The difference between these two effects is marginally significant (p = 0.097), suggesting that the intervention may have been more effective for students with higher baseline preparation.

5.3. Predictors of Engagement

So far, we have examined the overall effect of assignment to the virtual course. However, engagement among students assigned to the treatment varied substantially in intensity. Using administrative data from AoPS, we are able to track course engagement with a high degree of precision.

Figure 2 displays the distribution of engagement, measured in months, among students who were offered access to the course.[18] Average engagement in this group was approximately one month. However, the distribution is sharply bimodal: roughly half of students engaged minimally (one week or less), while about 15 percent remained active for the full course duration, with the remaining students exhibiting intermediate levels of participation. This pattern indicates substantial heterogeneity in how students respond once access is provided, underscoring the importance of understanding engagement as a distinct margin of response.

Table 4 examines the correlates of engagement intensity, measured as the number of months of active participation. The strongest and most consistent predictor of engagement

[18] Figure A.3 provides an alternative visualization.

is the selection-test score. A one–standard deviation increase in the selection-test score is associated with 0.27 additional months of engagement in the bivariate specification and 0.19 additional months when demographic and socioeconomic controls are included. This pattern indicates that students with stronger prior preparation were more likely to persist with the course.

Among the additional covariates, being female is associated with lower measured engagement (–0.43 months), a difference that is statistically significant at the 5 percent level. Other baseline characteristics—including age, English fluency, prior competition experience, socioeconomic background, parental education, and country-level IMO performance—are not significantly related to engagement.

Overall, the explanatory power of these regressions is modest: the full specification explains approximately 9 percent of the variation in engagement. This suggests that, even among a highly selected group of students, predicting ex ante who will remain actively engaged with the course is difficult.

5.4. Instrumental Variables Estimates

The experimental design furnishes a valid instrument for addressing the endogeneity of course engagement. While the intention-to-treat (ITT) specification identifies the effect of *assignment* to the course, we are also interested in recovering the causal effect of *actual engagement* on outcomes. Direct OLS estimation is inconsistent because engagement is endogenous: students who choose to engage more intensively may differ systematically in both observed and unobserved characteristics correlated with outcomes.

We therefore implement a two-stage least squares (2SLS) strategy, using randomized assignment to the course as an instrument for engagement. The first stage estimates the causal effect of assignment on engagement:

$$L_i = \pi \, Treat_i + \alpha_i + \boldsymbol{\delta}^\top \mathbf{X}_i + u_i, \tag{2}$$

where L_i denotes engagement, $Treat_i$ is the randomized offer, and α_i and $\mathbf{X}_i$ represent fixed effects and baseline covariates, respectively.

The second stage then regresses outcomes on the fitted values of engagement:

$$Y_i = \beta \, \widehat{L}_i + \alpha_i + \boldsymbol{\gamma}^\top \mathbf{X}_i + \varepsilon_i. \tag{3}$$

Under the standard IV assumptions—exclusion, relevance, monotonicity, and random assignment ensuring independence—β identifies the Local Average Treatment Effect (LATE):

the causal effect of engagement for compliers, i.e. students whose engagement status is shifted by the randomized offer.

Table 5 presents the two-stage least-squares (2SLS) estimates of the effect of course engagement on final exam performance. Across specifications, each additional month of course engagement increases standardized final-exam scores by 0.14–0.16 standard deviations. These estimates are larger than the ITT effects reported in Table 2, consistent with the fact that not all students assigned to the virtual course engaged with the material.

The specification including strata fixed effects (column 2) yields an estimated effect of 0.157 standard deviations per month of engagement, significant at the 5 percent level. Adding the full set of baseline controls (column 3) produces a very similar estimate of 0.165 standard deviations, significant at the 1 percent level. The stability of these estimates across specifications reinforces their robustness.

Because engagement is instrumented with random assignment to the virtual course, these estimates capture the causal effect of engagement for compliers—students who would engage with the course if offered access but would not otherwise. The strong first-stage relationship (approximately one additional month of engagement per assignment; F-statistics above 160) provides confidence in the relevance of the instrument.

The magnitude of these effects is meaningful. The course ran for 18 weeks—approximately 4.5 months—implying that a student who remained engaged throughout would improve their final exam score by about 0.6–0.7 standard deviations relative to no engagement. As reported in Section 5.3, about 15 percent of students assigned to the course achieved this level of engagement. Even partial engagement yields nontrivial benefits: two months of active participation correspond to gains of roughly 0.3 standard deviations.

5.5. Missing Final Exam Scores and Robustness

Approximately 40 percent of participants did not take the final exam. In the main analysis (Tables 2–5), we code missing final-exam scores as zero, interpreting nonparticipation as an absence of demonstrated mastery.[19] This coding choice avoids imposing model-based assumptions about unobserved outcomes, but if attrition differs systematically between treatment and control groups, the estimated effects could be biased. We therefore examine both the determinants of missingness and the robustness of our results to alternative assumptions about missing outcomes.

[19] A related approach is used by Angrist, Bettinger, and Kremer (2006), who address selective test-taking by constructing outcome measures that assign non-test-takers a score of zero. By assigning the same outcome to treated and control students who do not take the exam, this coding implicitly assumes that the treatment has no effect among non-test-takers. To the extent that some treated non-test-takers benefited from the course, the resulting estimates should be interpreted as a lower bound on the treatment effect.

Table A1 shows that assignment to the virtual course does not predict missingness: students assigned to the course are no more or less likely to take the final exam than students in the control group. The only robust predictor of nonparticipation is baseline performance—students with lower selection-test scores are significantly less likely to complete the final exam. This pattern suggests that missingness primarily reflects differences in ability rather than selective attrition induced by treatment. Moreover, there is no evidence that treatment assignment interacts systematically with baseline characteristics in predicting exam participation.

Table A2 examines the relationship between engagement with the assigned learning materials and final exam participation within each treatment arm. Among students assigned to the virtual course, each additional month of active engagement is associated with a 12 percentage point reduction in the probability of missing the final exam ($p<0.01$). Similarly, among control-group students, greater engagement with the ebook is strongly associated with lower rates of exam non-participation: a 10 percentage point increase in the share of the book read corresponds to a 5 percentage point reduction in the probability of missing the exam ($p<0.01$).

Students with weaker baseline preparation are disproportionately likely to be missing final exam scores, while higher-performing students face strong incentives to complete the assessment. Engagement patterns further support this interpretation: as shown in Table A2, non-test-takers exhibit substantially lower engagement with the learning materials in both treatment arms. Students who did not take the exam are thus characterized by both weaker baseline preparation and limited engagement with the course or ebook. Taken together, baseline performance and engagement patterns indicate that missing outcomes are unlikely to correspond to a meaningful share of high-achieving students. Nevertheless, we assess the sensitivity of our results to alternative treatments of missing data.

Table A3 reports intention-to-treat estimates under a range of assumptions about missing final-exam scores.[20] In addition to the baseline coding of missing scores as zero, we consider three percentile-based imputations that assign missing outcomes the 25th, 50th (median), and 75th percentiles of the control-group score distribution, as well as a model-based imputation using predictions from baseline covariates. As the imputed value assigned to missing observations increases, the estimated treatment effect declines monotonically, as expected mechanically. Importantly, the estimated effect remains positive and statistically significant

[20] A natural alternative approach is to report Lee (2009) bounds for selective observation. In our setting, however, such bounds are not especially informative. Assignment has essentially no effect on exam participation (Table A1), so Lee-type trimming is driven by small differences in observed participation rates and is sensitive to sampling variation. With approximately 40 percent of outcomes missing, the resulting bounds are wide and, in finite samples, can easily include negative values, limiting their ability to sharpen inference. In addition, implementing trimming in a manner consistent with the stratified randomization would require applying it within strata, which further increases variability given cell sizes. For these reasons, we do not report Lee bounds and focus on full-sample intention-to-treat estimates and the sensitivity analyses in Table A3, which directly assess robustness to economically meaningful alternative assumptions about missing outcomes.

under all plausible imputations, including the 25th- and 50th-percentile assignments and the model-based imputation. Only under the extreme assumption that missing students would have performed at the 75th percentile of the control distribution does the estimate become marginally insignificant.

Overall, these results indicate that the main findings are not sensitive to the treatment of missing outcomes. The persistence of positive and substantively meaningful effects under a wide range of alternative assumptions—together with evidence that attrition is related to baseline ability rather than treatment status—supports the robustness of the main conclusions.

5.6. International Mathematical Olympiad Outcomes

We also examine whether access to the virtual course is associated with subsequent performance in high-level mathematics competitions. Table 6 reports effects on participation and outcomes in the next International Mathematical Olympiad (IMO)[21], widely regarded as the most prestigious and difficult mathematics competition for secondary school students. Only 39 students in our sample went on to compete in the next IMO, so these results should be interpreted cautiously.

We find no effect on the likelihood of participating in the IMO (Column 1). The control mean is 0.071, and the estimated coefficient is close to zero and statistically insignificant. Because IMO participation is a rare binary outcome, the experiment has limited power to detect modest effects on this margin. Although gains in ability might be expected to influence selection outcomes, national selection exams may emphasize topics other than combinatorics or may not finely rank students at the top of the distribution, reducing the extent to which such gains affect the likelihood of progressing to the IMO.

Conditional on participating, we detect a statistically significant increase in performance on the combinatorics problem (Column 2). The point estimate of 2.83 points ($p < 0.05$) represents a substantial share of the control mean of 5.96. Given the difficulty of IMO items—typically scored on a 0–7 scale and designed to differentiate among extremely well-prepared students—this improvement suggests that the course content translated into enhanced performance in settings requiring advanced combinatorial reasoning.

By contrast, we find no evidence of effects on non-combinatorics topics such as geometry, number theory, and algebra (Column 3). The point estimate of -0.37 is small relative to the control mean of 6.91 and statistically insignificant, consistent with the domain-specific nature

[21] The timing of the intervention implies different IMO cycles across cohorts: Cohort 1 (treated in Spring 2023) could in principle be selected for IMO 2023; Cohorts 2 and 3 (treated in Fall 2023 and Spring 2024) correspond to IMO 2024; and Cohort 4 (treated in Fall 2024) corresponds to IMO 2025. The regressions include cohort fixed effects to absorb these timing differences.

of the instruction.

Overall, although based on a small sample, these results suggest that the gains documented in the previous subsections extend to external assessments of advanced combinatorial problem-solving, while remaining limited to the domain targeted by the course. This provides rare experimental evidence that online instruction can potentially improve performance at the frontier of secondary-school mathematics.

6. Conclusion

This paper asks whether expanding access to advanced training can accelerate learning among individuals with exceptional ability, and what limits the realization of large gains when access is provided. The question matters because exceptional talent drives a disproportionate share of innovation and scientific progress, yet many individuals with high ability may never realize their potential. While access to advanced instruction has traditionally been concentrated in a small number of elite institutions, scalable technologies offer a way to relax this constraint. Yet little causal evidence exists on how frontier learning responds when access is expanded, or on which constraints become binding once it is.

Using a randomized expansion of access to advanced problem-solving instruction for 620 highly selected students across 44 countries, we document four main findings.

First, large learning gains at the frontier are achievable through scalable online instruction. Assignment to high-quality remote instruction raises performance on an advanced mathematics exam by 0.16 standard deviations, and for students who remain engaged throughout the 18-week course, learning gains are large—on the order of 0.66 standard deviations. These gains extend beyond the study's assessment: among students who later compete in the International Mathematical Olympiad, those assigned to the course perform substantially better on the relevant combinatorics problems—an increase of roughly 50 percent relative to the control mean. Taken together, these results suggest that realizing exceptional ability does not *inherently* require dense, in-person elite environments.

Second, when access to advanced training is expanded, sustained engagement—rather than instructional quality alone—emerges as a binding constraint. Although access raises learning on average, aggregate gains fall far short of what is achievable because many students disengage. Roughly half of students assigned to the course participated minimally, while only 15 percent remained fully engaged throughout the 18-week course. This occurs despite the fact that participants are drawn from a highly selected population with demonstrated ability, strong intrinsic motivation, and explicit financial incentives. These patterns suggest that a key limitation may be sustaining effort once access is provided.

Third, sustained engagement is neither easily predicted nor readily induced. Observable baseline characteristics explain less than 10 percent of the variation in participation, while engagement during the program is highly informative about learning returns. This combination—large potential gains, weak ex ante predictability, and strong ex post signals—raises questions about static approaches to identifying and investing in exceptional ability.

Fourth, expanding access to advanced training does not mechanically equalize opportunity at the frontier. Learning gains are no larger for students outside capital cities or from countries with weaker traditions of advanced mathematics training. If anything, effects are larger for students with stronger existing preparation. These patterns suggest that expanded access to advanced training may complement existing systems of preparation and support, rather than substituting for them.

These findings have implications for programs aimed at developing advanced talent. Expanding access can relax an important constraint, but it does not by itself ensure sustained learning. Because returns to engagement are potentially large yet persistence is difficult to predict in advance, programs that allocate resources primarily based on baseline characteristics risk both overinvesting in participants who disengage and underinvesting in those whose engagement would materialize with continued exposure. One alternative is to provide broad initial access, observe engagement and progress, and adjust investment accordingly.[22] The heterogeneity in our results also suggests that scalable instruction is likely to be more effective when combined with conditions that support sustained engagement.[23]

These results also speak to current optimism surrounding artificial intelligence and personalized learning technologies. While such tools may further reduce the cost of delivering advanced training, our findings suggest that lowering instructional costs alone does not address the persistence challenge. Even among highly motivated, high-ability students with free access to excellent content, sustained engagement remains difficult to achieve. At the same time, whether current engagement rates are discouraging or promising depends partly on scale and cost. Even 15 percent full engagement, applied to a global population of high-ability students at low marginal cost, would represent a substantial expansion of realized potential relative to a world in which most such students lack access entirely. But realizing this potential at scale will depend not only on technological advances that lower costs but on complementary structures that support persistence.[24]

[22] This logic mirrors approaches in other domains where initial promise poorly predicts subsequent performance. A well-known example comes from football manager Brian Clough, who emphasized continuous reassessment—reinvesting in players who demonstrated sustained effort rather than relying on initial reputation alone.

[23] We cannot rule out that alternative program designs—such as different incentive structures, gamification, shorter modules, or more intensive accountability—might improve persistence. But it is notable that engagement remained limited even in a setting designed to be favorable: highly selected and motivated students, directly relevant content, a leading provider, and meaningful financial incentives. This suggests that sustaining engagement on demanding material may be even more challenging than commonly assumed.

[24] This is consistent with the possibility that elite assessments like the International Mathematical Olympiad select on

While our study focuses on students at the extreme upper tail of the ability distribution, some findings may have broader relevance. First, if persistence is difficult to sustain even among highly motivated, strongly incentivized students, it is likely to be at least as challenging in less selected populations. Second, the finding that online instruction appears to complement rather than substitute for local support structures may apply more generally: students at any level may benefit more from online resources when they have access to mentors, peers, or supportive environments. Third, the logic of dynamic allocation—providing broad initial access and adjusting investment based on revealed engagement—may be relevant for talent development beyond the settings studied here. To the extent that what sustains engagement varies across individuals, this suggests value in experimentation with different approaches rather than reliance on fixed program designs.

Overall, the evidence points to a cautiously optimistic conclusion. Learning at the frontier can be accelerated through scalable instruction, and large gains are potentially attainable even at the extreme upper tail of the ability distribution. But expanding access is only one step. Even when advanced training is available, many individuals with exceptional ability may never realize their potential because sustained engagement is difficult to maintain. Understanding—and designing for—the constraints that emerge after access is relaxed is therefore central to developing exceptional ability at scale. Viewed through this lens, effective talent development systems must address both sides of the equation: the supply of advanced training, which scalable technologies can increasingly provide at lower cost, and the demand-side forces—motivation, persistence, and supportive environments—that determine whether individuals ultimately benefit from that training.

persistence as much as on ability. If so, expanding access to advanced training shifts the bottleneck from identifying who has ability to identifying—or inducing—who will persist.

References

[1] Adajar, P. M., Duflo, E., Ellison, G., Ellison, S. F., & Kannan, H. (2025). Identifying and Nurturing Math Talent: Evidence from Tamil Nadu. *Unpublished Manuscript*, Massachusetts Institute of Technology.

[2] Agarwal, R., & Gaule, P. (2020). Invisible Geniuses: Could the Knowledge Frontier Advance Faster? *American Economic Review*, 110(3), 1068–1105.

[3] Agarwal, R., Ganguli, I., Gaule, P., & Smith, G. (2023). Why U.S. immigration matters for the global advancement of science. *Research Policy*, 52(1), 104659.

[4] Agarwal, R., Gaule, P., & Jiang, Y. (2025). Finding Young Einsteins: Olympiads and STEM Talent Discovery. Mimeo, University of Bristol.

[5] Angrist, J. D., Bettinger, E., & Kremer, M. (2006). Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia. *American Economic Review*, 96(3), 847–862.

[6] Bai, Y., Hsieh, M. H., Liu, J., & Tabord-Meehan, M. (2024). Revisiting the Analysis of Matched-Pair and Stratified Experiments in the Presence of Attrition. *Journal of Applied Econometrics*, 39(2), 256–268.

[7] Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying Education: Evidence from Two Randomized Experiments in India. *The Quarterly Journal of Economics*, 122(3), 1235–1264.

[8] Bell, A., Chetty, R., Jaravel, X., Petkova, N., & Van Reenen, J. (2019). Who Becomes an Inventor in America? The Importance of Exposure to Innovation. *The Quarterly Journal of Economics*, 134(2), 647–713.

[9] Beheshti, R. (2018). Maryam Mirzakhani (1977–2017). *Notices of the American Mathematical Society*, 65(10), 1221–1240.

[10] Bettinger, E. P., Fox, L., Loeb, S., & Taylor, E. S. (2017). Virtual Classrooms: How Online College Courses Affect Student Success. *American Economic Review*, 107(9), 2855–2875.

[11] Bruhn, M., & McKenzie, D. (2009). In Pursuit of Balance: Randomization in Practice in Development Field Experiments. *American Economic Journal: Applied Economics*, 1(4), 200–232.

[12] Cacault, M. P., Hildebrand, C., Laurent-Lucchetti, J., & Pellizzari, M. (2021). Distance Learning in Higher Education: Evidence from a Randomized Experiment. *Journal of the European Economic Association*, 19(4), 2322–2372.

[13] Card, D., & Giuliano, L. (2014). Does Gifted Education Work? For Which Students? Technical Report w20453, National Bureau of Economic Research, Cambridge, MA.

[14] Card, D., & Giuliano, L. (2016). Can Tracking Raise the Test Scores of High-Ability Minority Students? *American Economic Review*, 106(10), 2783–2816.

[15] de Losada, M. F., & Taylor, P. J. (2022). Perspectives on Mathematics Competitions and Their Relationship with Mathematics Education. *ZDM–Mathematics Education*, 54(5), 941–959.

[16] Duflo, E., Dupas, P., & Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, 101(5), 1739–1774.

[17] Duflo, E., Glennerster, R., & Kremer, M. (2007). Using Randomization in Development Economics Research: A Toolkit. In T. P. Schultz & J. A. Strauss (Eds.), *Handbook of Development Economics* (Vol. 4, pp. 3895–3962). Elsevier.

[18] Figlio, D., Rush, M., & Yin, L. (2013). Is It Live or Is It Internet? Experimental Estimates of the Effects of Online Instruction on Student Learning. *Journal of Labor Economics*, 31(4), 763–784.

[19] Gong, J., Liu, T. X., & Tang, J. (2021). How Monetary Incentives Improve Outcomes in MOOCs: Evidence from a Field Experiment. *Journal of Economic Behavior & Organization*, 190, 905–921.

[20] Hsieh, C.-T., Hurst, E., Jones, C. I., & Klenow, P. J. (2019). The Allocation of Talent and U.S. Economic Growth. *Econometrica*, 87(5), 1439–1474.

[21] Kofoed, M. S., Gebhart, L., Gilmore, D., & Moschitto, R. (2024). Zooming to Class? Experimental Evidence on College Students' Online Learning During Covid-19. *American Economic Review: Insights*, 6(3), 324–340.

[22] Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *Review of Economic Studies*, 76(3), 1071–1102.

[23] Lichand, G., Doria, C. A., Leal-Neto, O., & Fernandes, J. P. C. (2022). The Impacts of Remote Learning in Secondary Education During the Pandemic in Brazil. *Nature Human Behaviour*, 6(8), 1079–1086.

[24] McKenzie, D. (2022). Why I Am Now More Cautious About Using or Recommending Matched Pair Randomization and Like Matched Quadruplets Instead. World Bank Development Impact Blog.

[25] Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India. *American Economic Review*, 109(4), 1426–1460.

Tables

Table 1: Summary Statistics and Balance Tests

	Pooled	Control (0)			Treatment (1)			p-value
	Mean	Mean	SD	N	Mean	SD	N	of T − C
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	Difference (8)
Panel A. Demographics								
Female	0.231	0.226	0.423	310	0.235	0.423	310	0.774
Age	16.518	16.632	1.144	310	16.403	1.127	310	0.013
Panel B. Online Learning Prerequisites								
English fluency= beginner	0.075	0.068	0.246	310	0.081	0.264	310	0.527
English fluency= intermediate	0.352	0.368	0.475	310	0.335	0.475	310	0.364
English fluency= advanced	0.575	0.565	0.493	310	0.584	0.493	310	0.341
Internet at home	0.957	0.965	0.194	310	0.948	0.229	310	0.295
Family owns computer	0.902	0.916	0.282	310	0.887	0.317	310	0.205
Panel C. Socio-Economic Background								
Parent has university degree	0.763	0.784	0.405	310	0.742	0.440	310	0.178
Family owns car	0.726	0.735	0.440	310	0.716	0.458	310	0.562
Sleeps in own room	0.554	0.568	0.493	310	0.539	0.493	310	0.426
Has traveled internationally	0.570	0.613	0.493	310	0.526	0.493	310	0.019
Panel D. Preparation and Context								
Country IMO rank (lower = better)	57.075	57.855	31.517	310	56.294	31.164	310	0.756
Lives in capital city	0.453	0.474	0.493	310	0.432	0.493	310	0.267
Competition experience	4.719	4.790	2.254	310	4.648	2.095	310	0.151
Selection test score (standardized)	0.000	-0.009	0.951	310	0.009	0.951	310	0.971

Notes: F-test of joint significance of all differences between treatment and control groups: F = 1.368 (N = 620), excluding one English fluency category to avoid multicollinearity.

Table 2: Effect of the Virtual Course Offer on Final Exam Score

	Final exam score (standardized)		
	(1)	(2)	(3)
Assigned to virtual course	0.150*	0.165**	0.173**
	(0.080)	(0.073)	(0.073)
Controls	No	No	Yes
Strata FE	No	Yes	Yes
Observations	620	620	620
R^2	0.006	0.386	0.402

Notes: This table reports intention-to-treat estimates of the effect of assignment to the virtual course on standardized final-exam scores (mean = 0, SD = 1). Students who did not take the final exam are coded as having an outcome of zero, treating exam non-participation as an informative outcome rather than as missing test scores. Column 1 includes no controls, Column 2 adds strata (matched-quadruplet) fixed effects, and Column 3 includes baseline covariates (country IMO rank, years of competition experience, gender, age, capital-city indicator, parental education, socioeconomic index, and English fluency) in addition to the strata fixed effects. Robust (Eicker–Huber–White) standard errors in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Heterogeneous Effects of the Virtual Course Offer

	Final exam score (standardized)					
	Math tradition		Location		Selection score	
	(1) strong	(2) weaker	(3) capital	(4) outside	(5) high	(6) low
Assigned to virtual course	0.196	0.093	0.269**	0.048	0.254**	0.006
	(0.127)	(0.100)	(0.119)	(0.109)	(0.118)	(0.093)
Observations	300	320	281	339	337	283
R^2	0.008	0.003	0.018	0.001	0.014	0.000
p-value (equality across subgroups)	0.521		0.172		0.097	

Notes: This table reports intention-to-treat estimates by subgroup. The dependent variable is the standardized final-exam score; students who did not take the final exam are assigned a score of zero. Specifications exclude strata fixed effects as subgroup analysis would result in too few observations per stratum, leading to substantial sample attrition. Each column reports a separate regression without additional controls or strata fixed effects. Columns 1–2 split the sample by whether students are from countries with stronger versus weaker mathematical olympiad traditions, based on whether a country's IMO rank is better (below-median rank) or worse (above-median rank). Columns 3–4 split the sample by whether students are based in their country's capital city. Columns 5–6 split the sample by whether students scored above or below the median on the selection test. The reported p-values test equality of treatment effects across each pair of subgroups. Robust standard errors in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Predictors of Engagement with the Virtual Course

	Months of course engagement	
	(1)	(2)
Selection test score (standardized)	0.272***	0.193**
	(0.085)	(0.089)
Country IMO rank (lower=better)		-0.004
		(0.003)
Competition experience		0.039
		(0.044)
Female		-0.428**
		(0.195)
Age		-0.061
		(0.072)
Lives in capital city		-0.187
		(0.164)
Socioeconomic index		0.107
		(0.069)
Parent has university degree		0.109
		(0.190)
English fluency = beginner		0.081
		(0.312)
English fluency = intermediate		-0.092
		(0.183)
Observations	310	310
R^2	0.032	0.086
Dependent variable mean	1.06	1.06

Notes: The sample is restricted to students assigned to the virtual course. The dependent variable is the number of months of active engagement with the virtual course, derived from AoPS participation data. Column 1 includes only the selection test score; Column 2 adds demographic and socioeconomic controls. Robust standard errors in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Effect of Months of Course Engagement on Final Exam Score

	Panel A: Second Stage final exam score (standardized)		
	(1)	(2)	(3)
Months of course engagement	0.142* (0.074)	0.157** (0.059)	0.165*** (0.057)
	Panel B: First Stage months of course engagement		
Random assignment to course offer	1.058*** (0.082)	1.053*** (0.076)	1.062*** (0.078)
Controls	No	No	Yes
Strata FE	No	Yes	Yes
Observations	620	620	620
First-stage F-stat	168.3	191.2	186.5

Notes: Panel A reports two-stage least-squares estimates of the effect of months of course engagement on standardized final-exam scores, instrumenting engagement with assignment to the virtual course. Panel B reports the first-stage relationship between assignment and engagement. Students who did not take the final exam are coded as having an outcome of zero, treating exam non-participation as an informative outcome rather than as missing test scores. Column 1 includes no controls, Column 2 adds strata fixed effects, and Column 3 includes both baseline covariates and strata fixed effects. Robust standard errors in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Effect of the Virtual Course Offer on Subsequent IMO Performance

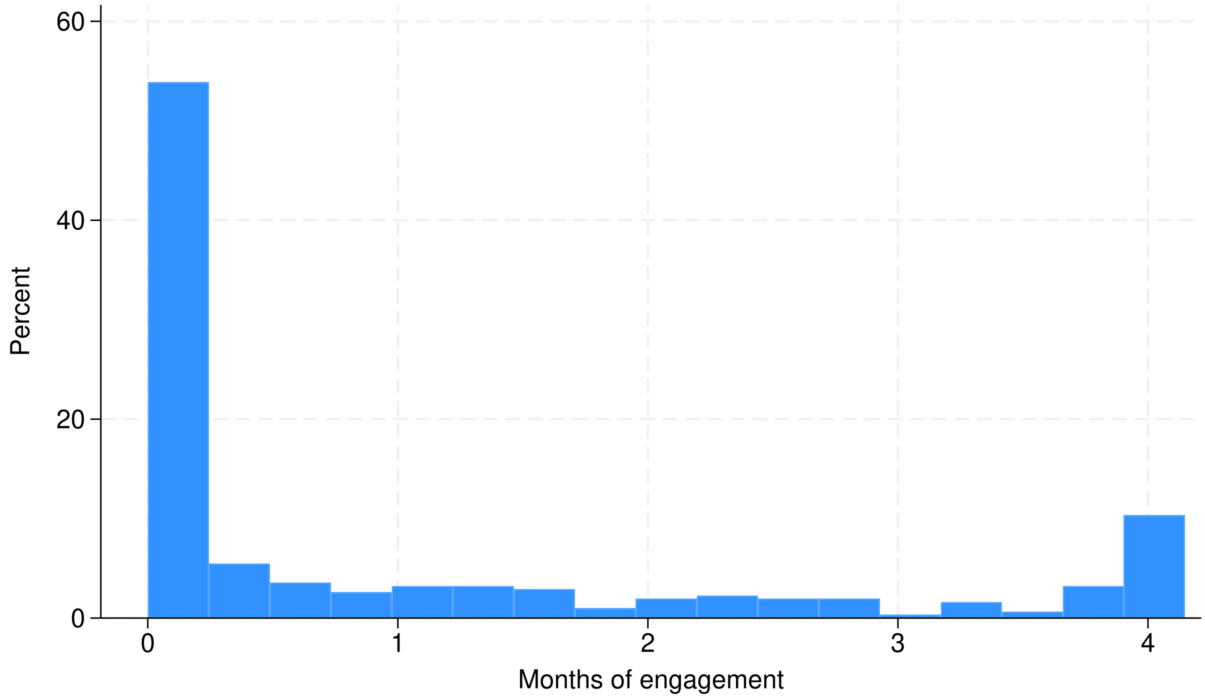
	Participated in next IMO	**Combinatorics score**	**Other topics score**
	(1)	(2)	(3)
Assigned to virtual course	-0.016	2.831**	−0.371
	(0.019)	(1.274)	(1.659)
Observations	620	39	39
Control mean	0.071	5.96	6.91

Notes: This table examines the effect of assignment to the virtual course on subsequent International Mathematical Olympiad (IMO) outcomes. Column 1 reports the effect on participation in the next IMO using the full sample ($N = 620$). Columns 2 and 3 report effects on performance in combinatorics and in other topics (geometry, number theory, and algebra), conditional on participating in the next IMO ($N = 39$). Scores range from 0 to 7 per problem and are aggregated within domains: the combinatorics score sums performance on the combinatorics problems, and the 'other topics' score sums performance on the four remaining problems. Robust standard errors in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure 1: Countries with participants in the study

Notes: This figure shows the countries represented in the study sample. Participants were primarily recruited through national mathematics olympiad organizations, which referred high-performing students. In a small number of cases, students contacted the research team independently.

Figure 2: Distribution of Student Engagement



Notes: This figure shows the distribution of course engagement among students assigned to the virtual course ($n = 310$). Engagement is based primarily on completion of graded problem sets (including 'engaged but struggled'), recorded weekly by AoPS.

Appendix A: Further Details on Mathematics Competitions

A.1. The International Mathematical Olympiad

The International Mathematical Olympiad (IMO) is the most prominent and selective mathematics competition for secondary school students worldwide. It has been held annually since 1959 and currently involves more than 100 participating countries. Each participating country may send a team of up to six students, typically selected through national competitions and training programs.

Each year, the IMO is hosted by a single country, with all participating teams traveling to a common location.[25] The competition is conducted simultaneously for all participants under uniform rules and time limits, ensuring comparability of performance across countries. Although students participate as members of national teams, all examination work is completed individually, and scores are assigned on an individual basis.

A defining feature of the IMO is that a completely new set of problems is created each year specifically for the competition. Problems are proposed by participating countries in the months leading up to the event and undergo an extensive, multi-stage vetting process overseen by the IMO Jury, which consists of official representatives (typically team leaders) from all participating nations. The Jury meets before the competition to review proposed problems, assess their difficulty and originality, and select the final six problems. This process is designed to ensure broad coverage of mathematical topics, appropriate difficulty calibration, and prevention of prior exposure—participants confront problems they have not seen before and cannot prepare for by studying a fixed pool of questions.

Problems span four core areas of elementary mathematics—algebra, number theory, combinatorics, and geometry—with the precise mix varying across years. Within each examination day, problems are typically ordered by difficulty, with the third problem on each day (Problems 3 and 6) generally considered the most challenging. Even among highly selected participants, it is common for students to solve only a subset of the problems fully. Partial credit is awarded for correct intermediate steps, but full solutions require clear logical exposition and rigorous argumentation.

Solutions are written by students in their preferred language and are translated and graded according to standardized marking schemes agreed upon by the Jury, ensuring consistency in evaluation across countries and languages.

[25] In exceptional circumstances—most notably during the COVID-19 pandemic—the IMO has been administered remotely, with coordinated local venues.

A.2. Interpreting Competition Performance

Performance in mathematics competitions reflects a specific set of advanced problem-solving skills. Competition problems emphasize abstraction, logical reasoning, and the construction of rigorous proofs, often requiring students to recognize deep structural patterns and solve unfamiliar problems under time constraints. Unlike standardized tests that reward speed and procedural fluency, IMO problems typically admit multiple solution approaches and place a premium on creativity and insight. A student who identifies an elegant solution path may solve a problem in under an hour, while another student—equally capable in other respects—may struggle for the full session without making progress.

At the same time, competition performance captures only a subset of mathematical ability. The problems emphasize "elementary" mathematics—topics that can be stated without advanced prerequisites—rather than the technical machinery typically taught at the university level. Success requires little in the way of sustained computation, applied problem solving, or familiarity with advanced theories such as abstract algebra or real analysis. Performance may also be influenced by prior exposure to competition-style training and familiarity with common problem types and solution techniques, which vary substantially across countries and institutional settings. Two students with similar underlying aptitude may perform quite differently depending on the quality and duration of their preparation.

A.3. Relationship to Later Outcomes

Participants in elite mathematics competitions are disproportionately represented among later contributors to mathematical research and related fields. Several prominent mathematicians, including Fields Medalists Terence Tao and Maryam Mirzakhani, participated in the IMO during their secondary school years. Such examples are illustrative rather than representative and do not, on their own, establish a causal relationship between competition participation and later outcomes.

Systematic evidence on this relationship is provided by recent work linking mathematics competition performance to long-run scientific contributions. Agarwal & Gaule 2020 construct a comprehensive dataset linking IMO participants to later publication records and document that former participants are highly overrepresented among frontier scientific contributors, particularly in mathematics, physics, and computer science. Subsequent work examines the relationship between competition performance and later outcomes in greater detail. Agarwal et al. (2023) show that IMO scores predict research productivity and impact even within the highly selected sample of competition participants, and that this relationship is not fully explained by institutional placement or country of origin. Agarwal, Gaule & Jiang (2025) document a strong relationship between International Mathematical Olympiad performance and subsequent receipt of major scientific prizes.

These studies emphasize that the observed relationships are descriptive rather than causal. IMO performance reflects a combination of underlying ability, sustained motivation, and access to high-quality training—all of which may independently predict later success. The competition itself may also provide valuable experiences, networks, and signals that shape subsequent opportunities. Disentangling these channels remains an important area for future research.

A.4. Limitations as a Talent Identification Mechanism

While mathematics competitions provide a valuable mechanism for identifying and developing mathematical talent, they have important limitations that should inform their interpretation and use.

First, participation is uneven across countries and socioeconomic groups. Students require not only underlying aptitude but also awareness of competition opportunities, access to preparatory resources, and sustained institutional support. In many countries, high-quality training is concentrated in major cities or available only through expensive private instruction, limiting participation among students from rural areas or lower-income families. Even where access is formally open, the time and resources required for sustained preparation may be prohibitive for students with other obligations.

Second, early identification through competition systems may disadvantage students who develop mathematical interests or abilities later in adolescence. Competition success often depends on years of accumulated training, and students who enter the pipeline late—however talented—may struggle to catch up. The emphasis on early selection may therefore overlook individuals whose potential would become apparent given more time or different opportunities.

Third, the individual and time-constrained nature of competition problems may underrepresent other dimensions of mathematical competence. Professional mathematical research often involves extended engagement with a single problem over weeks or months, collaboration with peers, and the application of specialized technical tools. These aspects of mathematical work are largely absent from competition settings, which reward independent problem solving under pressure.

Finally, competition performance may be shaped by factors beyond mathematical ability, including test-taking strategies, psychological resilience under pressure, and familiarity with particular problem formats. While these factors may correlate with later success, they are conceptually distinct from the underlying construct of mathematical talent that competitions aim to measure.

Taken together, these limitations suggest that mathematics competition outcomes should be interpreted as informative but incomplete indicators of advanced mathematical ability. They provide a standardized and internationally comparable measure of performance among highly selected students, while capturing one dimension of a broader and multifaceted concept of mathematical talent.

Appendix B: Study Design, Recruitment, and Implementation

This appendix provides detailed information on the design and implementation of the randomized controlled trial. It documents the recruitment of participants and partner organizations, the screening and baseline assessment procedures, the randomization protocol, the instructional intervention and control condition, the measurement of engagement and learning outcomes, and key operational aspects of implementation. These details complement the discussion in the main text and are intended to support transparency and reproducibility of the study design.

Appendix Figure A.1 presents a schematic overview of the experimental design, illustrating the sequence of recruitment, baseline screening, randomization, treatment assignment, and outcome measurement.

B.5. Participant Recruitment and Partner Organizations

Recruiting participants for the study posed a substantial practical challenge. The target population—students with exceptional mathematical ability—is small by definition and geographically dispersed. In addition, the study prioritized students from developing and middle-income countries, where access to high-quality training resources is more limited. As a result, recruitment required institutional collaboration rather than open enrollment.

We relied primarily on national mathematics olympiad organizations to identify and invite eligible students. Contact information for official country representatives is publicly available through the International Mathematical Olympiad (IMO), and team leaders and delegates convene annually at the IMO to form the IMO Jury. To assess feasibility, we initially contacted IMO representatives from a range of developing countries and established collaborations with a first group of national organizations. After the first cohort, one of the authors attended the IMO in person to meet additional team leaders and expand the set of participating countries. We also collaborated with several regional olympiads, including the Pan African Mathematical Olympiad, the East African Mathematical Olympiad, and the Central American and Caribbean Mathematical Olympiad (OMCC), inviting medalists from these competitions to participate.

Participating national organizations viewed the study as an opportunity to provide advanced training to their students and, in some cases, to strengthen their broader talent-development pipelines beyond the small group ultimately selected for national IMO teams. Because the pool of eligible students within each country was limited and onboarding new countries required time, the experiment was implemented over multiple cohorts spanning two academic years.

In most cases, partner organizations preferred not to share students' contact information directly. Instead, they circulated a recruitment flyer to selected students on our behalf. In some instances, organizations shared lists of student names without contact details, which we cross-checked against publicly available competition results. In other cases, eligibility relied on students' self-reported competition achievements. These constraints motivated the use of a standardized selection test to ensure that enrolled participants met the study's ability threshold.

Take-up rates varied substantially across countries and institutional contexts. For example, in Pakistan, where recruitment was conducted in close collaboration with COMSATS University Islamabad—the institution responsible for national olympiad selection and training—approximately 87 percent of students in the extended training cadre expressed interest in participating. In contrast, in Brazil, where recruitment occurred through OBMEP, the national public-school mathematics olympiad, take-up was closer to 15 percent.

Finally, some attrition occurred prior to randomization. A subset of students who initially expressed interest did not complete all registration steps, including signing the consent form, creating an AoPS account, or completing the baseline survey. Additional attrition occurred when some students did not attend the pre-randomization workshop or the selection test. These screening steps were implemented to ensure that the randomized sample consisted of students who were both eligible for and willing to engage with the intervention.

B.6. Recruitment Materials and Information Provided to Students

Recruitment was conducted using a standardized flyer circulated by partner organizations to selected students. The flyer described the study as the Search for Talented Ramanujans (STAR) program and framed participation as an opportunity to receive advanced training in mathematics, particularly in areas relevant to mathematics competitions, including counting, probability, discrete mathematics, and graph theory.

The flyer emphasized that participation in the program was free and fully remote. It highlighted several potential benefits of participation, including access to instructional materials from Art of Problem Solving (AoPS), the possibility of enrolling in an 18-week AoPS online course on counting and probability (with enrollment determined by random assignment), certificates recognizing participation and excellence, and the opportunity to compete for cash prizes. The flyer stated that cash prizes of up to USD 500 would be awarded to approximately one third of participants based on performance on a final test.

The flyer also outlined the main stages of the program, including an initial workshop, a selection test, the course period, and a final exam, and provided indicative dates for each stage. In addition, it noted that participation could help students prepare for mathematics competitions and strengthen their profiles for future studies in STEM fields. No guarantees of course enrollment or prize receipt were made; instead, the flyer explicitly stated that enrollment in the AoPS course would be determined by random assignment.

Overall, the recruitment materials presented the study as an opportunity for motivated, high-achieving students to engage in advanced mathematics training, while transparently communicating the role of randomization and performance-based incentives.

B.7. Pre-Randomization AoPS Workshop

Beginning with Cohort 2, we introduced a pre-randomization instructional workshop conducted by Art of Problem Solving (AoPS). The workshop was designed to serve three purposes: (i) to familiarize participants with the AoPS instructional format, (ii) to observe initial engagement with competition-style online instruction, and (iii) to provide all participants with access to high-quality mathematics instruction prior to randomization.

The workshop consisted of two live instructional sessions of three hours each, delivered over two consecutive weekends, and was accompanied by homework assignments. Instruction covered introductory topics in counting and probability, as well as a selection of problems drawn from other areas of elementary mathematics. The level of difficulty was lower than that of the main course but reflected the problem-solving style and interactive pedagogy characteristic of AoPS online instruction.

Participation in the workshop formed part of the screening process for progression to the randomized trial. Students who showed minimal engagement with the workshop—such as failing to attend sessions or complete assigned problems—were not invited to proceed to the randomization phase. This step was intended to ensure that participants entering the trial had a basic familiarity with the instructional format and a minimum level of engagement consistent with the demands of the full course.

In addition to its screening role, the workshop ensured that all students who advanced beyond the initial recruitment stage received some instructional benefit, regardless of subsequent treatment assignment. The workshop was implemented uniformly across participants within each cohort and was completed prior to randomization. No outcome measures from the workshop were used in the experimental analysis.

B.8. Baseline Ability Measures and Selection Tests

Baseline measures of mathematical ability were used to characterize the study population and to support stratified randomization and regression adjustment. The availability and form of these measures differ across cohorts due to the staged rollout of the study.

For cohorts 2–4, baseline mathematical ability was measured using a standardized selection test administered prior to randomization. The selection tests were designed to assess readiness for the course content and focused primarily on combinatorics and relevant prerequisites, particularly algebra. Each test consisted of nine non-routine problems to be solved within a two-hour time limit. The problems were developed specifically for the study by an expert mathematician, who also designed the final exam used to measure learning outcomes.

The selection test questions required short numerical or symbolic answers only; no written explanations or proofs were requested. Responses were graded automatically as correct or incorrect. Within each cohort, all participants took the selection test simultaneously under uniform conditions. Separate selection tests were developed for cohorts 2, 3, and 4 to prevent prior exposure to test items.

In addition to providing a baseline measure of ability, the selection test served as an eligibility screen. Students who did not take the test or who demonstrated insufficient mastery of prerequisite material—typically by solving fewer than three to four questions correctly—were not invited to proceed to the next stage of the study and did not enter the randomization. This screening step ensured that the intervention was targeted to students for whom the course content was appropriate.

Selection test scores were standardized to have mean zero and unit variance in the pooled sample. These standardized scores were used for stratification, balance assessment, and as baseline controls in the empirical analysis.

The first cohort served as a pilot and was recruited before the selection test was implemented. For this cohort, baseline ability was proxied using alternative measures when available. For participants from Mongolia, we use standardized national olympiad results. For participants from Malaysia, we use standardized scores from national training camp assessments. For a small number of participants from the Philippines, for whom no comparable baseline measure was available, we assign the sample median value. These proxy measures were standardized in the same way as the selection test scores to ensure comparability across cohorts.

B.9. Randomization Implementation

Randomization was conducted after completion of all pre-randomization screening steps, including informed consent, the baseline survey, and the baseline ability assessment (selection test or proxy measure, depending on cohort). Only students who completed these steps and met the eligibility criteria described above proceeded to randomization.

Randomization was implemented separately for each cohort. For each cohort, one of the authors constructed strata using baseline information available prior to treatment assignment. In cohorts 2–4, strata were formed based on standardized baseline ability measures, with country (and country IMO rank) used as a secondary stratification variable when needed to resolve ties. In the first cohort, which served as a pilot, stratification was implemented first by country and then by baseline ability measures, reflecting differences in available information at that stage of the study.

The target stratum size was four students (matched quadruplets). In most cases, this yielded strata of equal size. A small number of strata (seven in total across all cohorts) contained between two and six students due to the total number of eligible participants not being divisible by four. These strata were retained rather than dropped to preserve sample size.

Within each stratum, randomization was conducted at the individual level. Random numbers were generated using Stata, and students were ranked within each stratum based on these random draws. Treatment assignment was then determined by rank: in strata of four, the two students with the highest random numbers were assigned to the treatment group and the remaining two to the control group. In strata of other sizes, assignment was adjusted proportionally to maintain approximate balance between treatment and control within strata.

Treatment assignment was implemented in a single step for each cohort. All randomized students are included in the analysis according to their original assignment.

B.10. Treatment and Control Conditions

B.10.1 Course Contents and Instructional Scope (Treatment)

The virtual course used in the experiment was Intermediate Counting & Probability, an 18-week course offered by Art of Problem Solving (AoPS) and designed for high-performing middle- and high-school students. The course focuses on discrete mathematics, with particular emphasis on combinatorics and probability, and is intended for students with prior exposure to competition-style problem solving.

AoPS delivered the course specifically for students enrolled in the experiment. Aside from scheduling, however, the course was identical in content, structure, and instructional approach to the corresponding AoPS courses offered commercially to U.S.-based, fee-paying students. Instruction was provided by AoPS instructors following the standard curriculum, materials, and grading practices used in regular offerings.

The curriculum covers a broad range of core topics in combinatorics and related areas, including one-to-one correspondences, the principle of inclusion–exclusion, constructive counting methods, distributions, recursion, generating functions, conditional probability, and introductory graph theory. These topics are central to mathematics competitions and are typically not taught in depth in standard secondary-school curricula. Instruction emphasizes solving novel problems rather than mastering a fixed set of techniques, with the goal of developing flexible reasoning and proof-oriented thinking.

The course is structured around weekly 90-minute live sessions and regular homework assignments, corresponding to approximately 4–5 hours of work per week over the duration of the course. Live sessions were scheduled to accommodate a global student body: classes were held at 4:00 pm UK time for students based in Europe, Africa, and the Americas, and at 1:00 am UK time (morning local time) for students in Central and East Asia. Sessions took place on weekends (Saturdays or Sundays) to avoid conflicts with school schedules.

As part of the course, students were expected to participate in live sessions and to complete homework assignments, including short-answer challenge problems and written solutions requiring formal reasoning. These activities were part of the instructional process and were not used to assess learning outcomes in the study.

B.10.2 The Control Condition

Students assigned to the control group did not receive access to the AoPS online course or any live instruction. Instead, they were provided free access to an AoPS e-book covering the same subject matter as the course, along with a structured study plan indicating suggested readings and exercises. This design isolates the effect of structured, instructor-led online instruction—including live interaction, feedback, and pacing—from access to high-quality written materials alone.

Importantly, the control condition reflects a realistic alternative available to motivated students: self-directed study using widely used competition-preparation materials. At the same time, control students did not receive graded assignments, individualized feedback, or live problem-solving sessions. As shown below, independent enrollment in AoPS online courses during the study period was extremely rare, and no student—treated or control—purchased access to any AoPS course in combinatorics, ensuring a clear separation between treatment and control.

B.11. Control-Group Access to AoPS Courses

In principle, students assigned to the control group could have independently purchased access to AoPS online courses during the study period, potentially reducing the contrast between treatment and control. We are able to assess this possibility directly using platform-wide administrative data from AoPS, which record all course enrollments and activity for study participants.

Across all students participating in the experiment, only eight individuals purchased access to any AoPS online course during the study period. None of these purchases involved the Intermediate Counting and Probability course used in the intervention, and more generally, no student—treated or control—purchased access to any AoPS course in combinatorics during the study period. Enrollment in other AoPS courses was rare and limited in scope.

B.12. Final Exam Design, Administration, and Grading

Learning outcomes were measured using a dedicated final exam designed specifically for the study. The exam was created by an expert mathematician whom we hired for this purpose. New problems were developed separately for each cohort to prevent prior exposure and to ensure that the assessment was appropriately challenging for a highly selected population.

Each final exam consisted of six problems drawn from combinatorics and closely related topics. The exam included a mix of short-answer questions and problems requiring students to present full solutions and reasoning. For problems requiring written solutions, students were asked to photograph and upload their handwritten work. Participants had two hours to complete the exam.

Exams were administered online and simultaneously for all participants within each cohort. For the first two cohorts, exams were delivered using Cuttle, a platform commonly used for mathematics competitions, including by the Australian Mathematics Trust. Later cohorts used Moodle. In one cohort, exam times were staggered across two time windows—one for students located in Central and East Asia and one for students located elsewhere—to accommodate time zone differences; within each window, the exam was administered simultaneously.

All exams were graded by the same expert mathematician who designed the problems, using anonymized submissions. Grading was independently checked by a second individual with relevant expertise. An appeals process was in place, allowing students to contest specific grades or raise concerns about official solutions, including the validity of alternative solution approaches. This process mirrors standard procedures used in mathematics competitions.

The exams proved highly challenging. While the majority of students earned at least partial credit on one or more problems, no participant achieved the maximum possible score of 42 points in any cohort. Mean raw scores ranged from approximately 5 to 13 out of 42 across cohorts, despite the presence of several International Mathematical Olympiad medalists among the participants. These outcomes indicate that ceiling effects are unlikely to be a concern.

Ideally, exams would have been proctored. In-person proctoring was not feasible given the geographic dispersion of participants across many countries, and remote proctoring was not pursued due to legal and ethical constraints related to monitoring minors across jurisdictions. Students therefore completed the exams remotely and independently.

Several forms of cheating could, in principle, be considered: use of existing online solutions, assistance from experts not enrolled in the study, collaboration with other students, or use of artificial intelligence tools. The use of existing materials is unlikely, as all problems were newly designed and exams were administered simultaneously. To assess the feasibility of AI assistance, we tested the problems using standard AI tools and ruled out any that could be readily solved. While assistance from experts or peers cannot be fully ruled out, the short exam duration, the difficulty of the problems, the limited pool of individuals capable of providing meaningful help, and the geographic dispersion and anonymity of participants make systematic assistance unlikely. We also cross-checked submitted solutions for similarities across students.

To the extent that any inappropriate assistance occurred, it does not appear to have been widespread or effective, given the low overall scores. Moreover, there is no clear reason to expect such behavior to differ systematically between treatment and control groups, which faced identical exam conditions and incentives.

B.13. Disbursement of Performance-based Incentives

The study offered cash prizes to incentivize sustained engagement and participation in the final assessment. Implementing these payments across a large number of individual recipients in many countries posed substantial practical challenges, primarily due to financial regulations related to anti–money laundering, restrictions on cross-border transfers, and the fact that some recipients did not have access to formal banking services.

In the first cohort, cash prizes were distributed indirectly. Payments were made as grants to the relevant national mathematics olympiad organizations, which then disbursed funds to students locally. In several cases, these payments were made in conjunction with award ceremonies organized by the national organizations. While this approach facilitated compliance with local regulations, it relied on country-specific arrangements and was not scalable to the full international scope of the study.

For subsequent cohorts, we adopted a centralized payment solution using the B4B Payments system. This system provides prepaid debit cards that can be remotely loaded and used for electronic purchases. The cards impose restrictions on use, including prohibitions on cash withdrawal and the purchase of alcohol or other restricted goods. These constraints were consistent with the study's goal of providing direct incentives to students while ensuring appropriate use.

Initially, physical debit cards were mailed to recipients. However, international postal delivery proved unreliable in many locations, with frequent delays or losses. As a result, we transitioned to issuing digital debit cards, which could be delivered electronically and activated remotely. This change substantially improved reliability and reduced administrative burden.

Throughout the study, the intention was for prizes to be awarded directly to students rather than to their families. At the same time, given that many participants were minors, parental awareness and consent were required. Parents or guardians were asked to provide explicit consent for their children to receive cash prizes through a signed authorization form.

B.14. Secondary Outcome: International Mathematical Olympiad Participation and Performance

As a secondary outcome, we examine whether access to the virtual course is associated with subsequent participation and performance in the International Mathematical Olympiad (IMO). The IMO represents the highest level of secondary-school mathematics competition and provides an externally administered, high-stakes assessment of advanced problem-solving ability.

Timing and cohort alignment. Because the intervention was implemented over multiple cohorts, outcomes related to IMO participation and performance necessarily correspond to different IMO editions. For each cohort, we define the relevant outcome as participation and performance in the first IMO held after completion of the intervention period. Specifically, Cohort 1 (treated in Spring 2023) could in principle be selected for the IMO 2023; Cohorts 2 and 3 (treated in Fall 2023 and Spring 2024) correspond to the IMO 2024; and Cohort 4 (treated in Fall 2024) corresponds to the IMO 2025. All regressions therefore include cohort fixed effects, which absorb differences in IMO year, exposure windows, and cohort-specific selection environments.

Data construction and sample. Information on IMO participation and problem-level scores was collected from publicly available IMO records. Participation in the IMO is observed for all study participants, while performance outcomes are available only for those who advanced to the competition. Because only a small subset of participants competed in the IMO, estimates based on IMO performance are necessarily imprecise and should be interpreted cautiously.

Classification of problems by topic. To assess whether the intervention affected performance in areas most closely aligned with the course content, we distinguish between combinatorics and non-combinatorics problems. For each IMO edition, problems were classified based on their primary mathematical content following standard topic conventions used in mathematics competitions. Combinatorics problems are those whose solutions rely primarily on counting arguments, discrete structures, or combinatorial constructions.

For the IMO 2023, problems 5 and 6 are classified as combinatorics; for the IMO 2024, problems 1, 3, and 5; and for the IMO 2025, problems 1, 5, and 6. For each participant, combinatorics performance is measured as the sum of scores on the combinatorics-designated problems for the relevant IMO year. Performance in other topics is defined analogously as the sum of scores on the remaining problems, which cover algebra, geometry, and number theory.

Interpretation. IMO outcomes provide a demanding and externally validated measure of performance among students at the extreme upper tail of the mathematics ability distribution. At the same time, participation in the IMO is determined by country-specific selection procedures that are outside the scope of the intervention, and performance estimates are based on a small and highly selected subsample. The IMO results should therefore be viewed as complementary to the primary experimental outcomes rather than as stand-alone evidence.

Appendix C: Supplementary Tables and Figures

Figure A.1: Randomized Controlled Trial Design

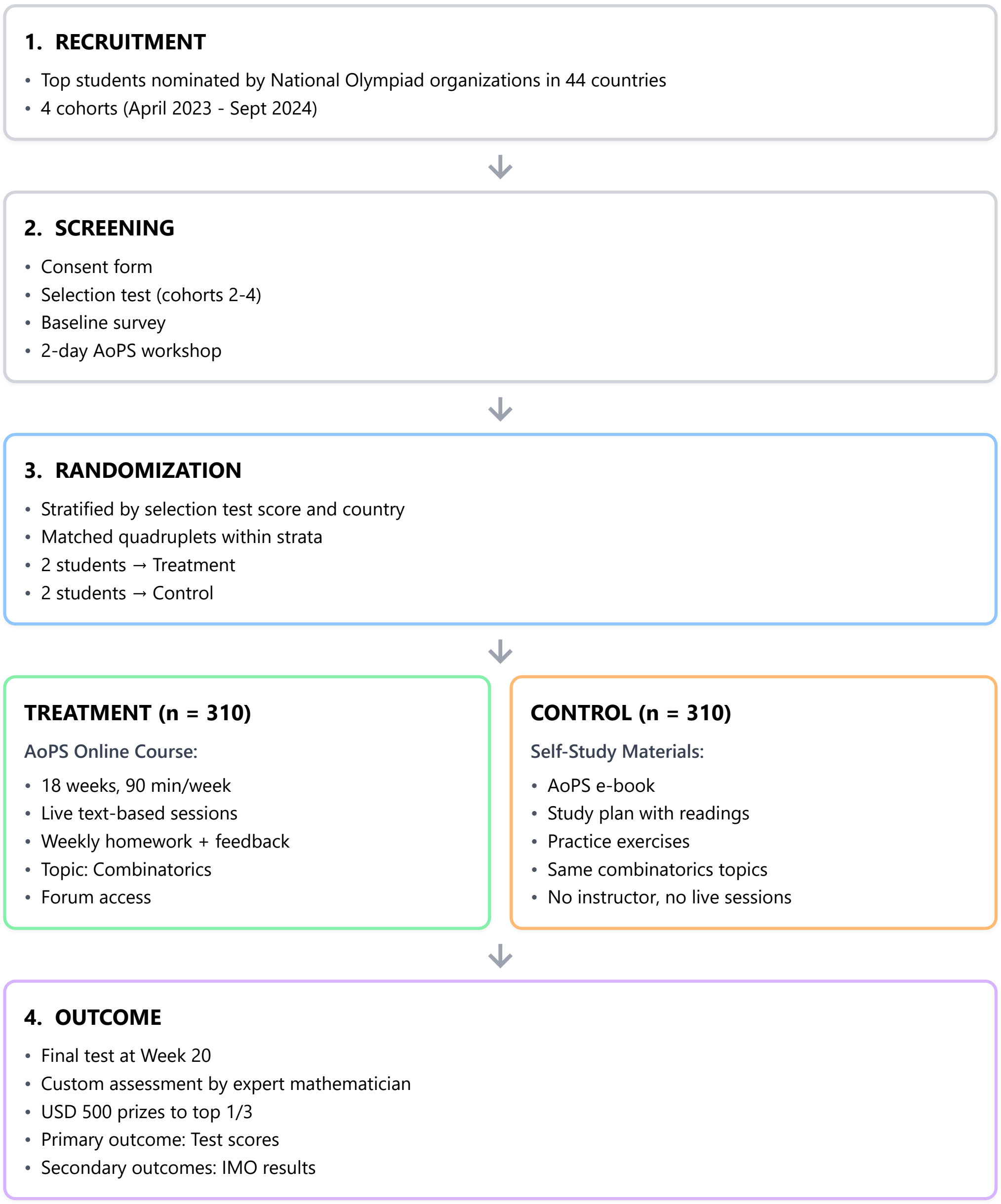
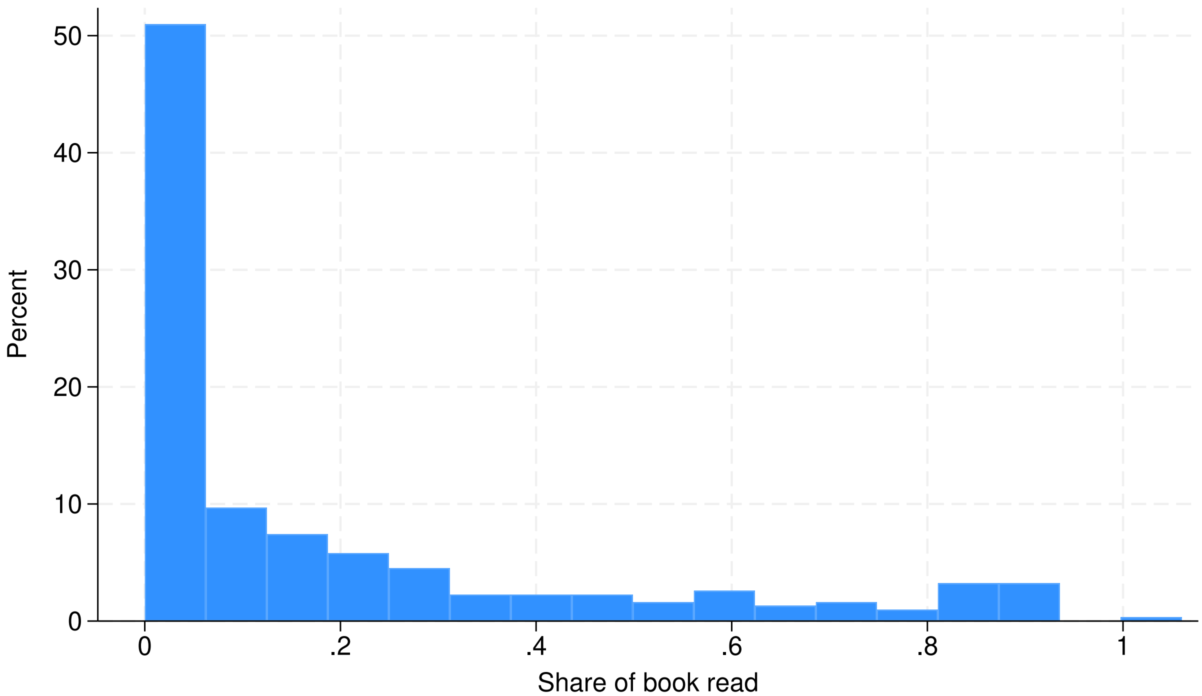
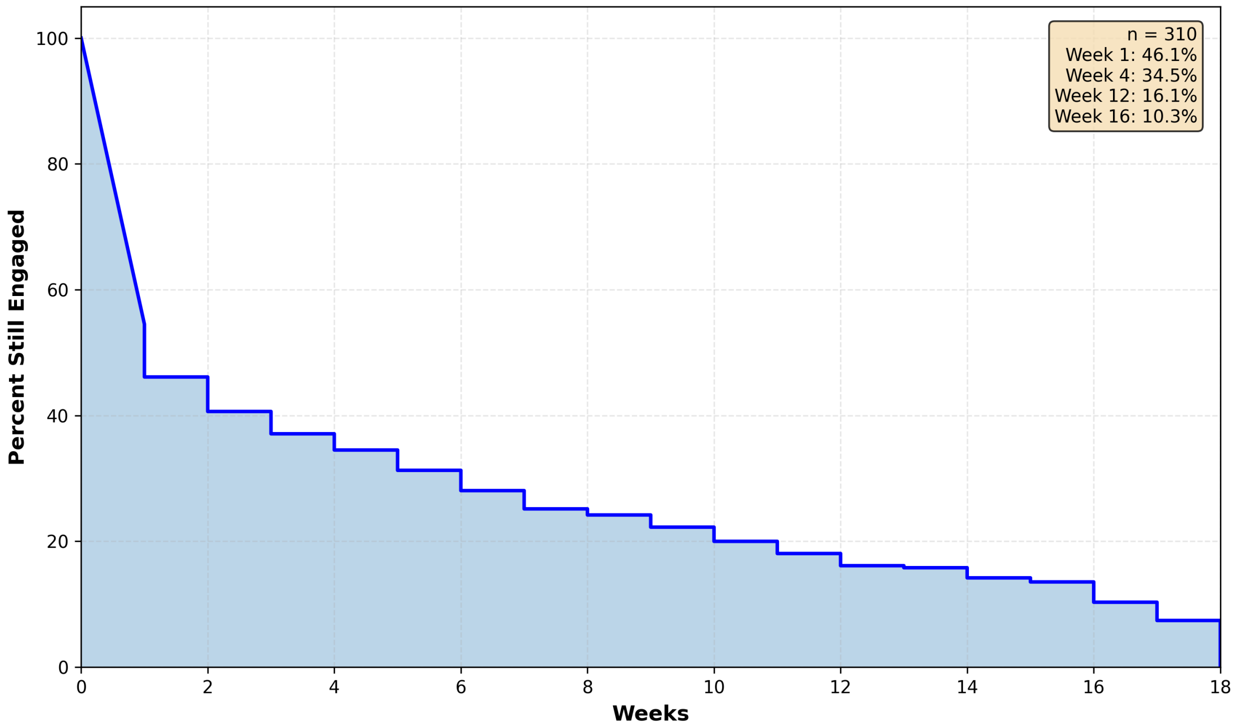


Figure A.2: Distribution of Ebook Engagement in the Control Group



Notes: Control group students (n=310) received access to an AoPS ebook covering the same content as the course. As an interactive online document (rather than a static PDF), the ebook allows us to track reading behavior. This figure shows the distribution of engagement, measured as the share of fragments read.

Figure A.3: Student Course Engagement Over Time



Notes: This figure plots persistence in course engagement among students assigned to the virtual course (n=310). Engagement is based primarily on completion of graded problem sets (including ’engaged but struggled’), recorded weekly by AoPS. The curve shows the share of students remaining engaged over time.

Table A1: Predictors of Final Exam Completion

	Did not take final exam		
	(1)	(2)	(3)
Assigned to virtual course	-0.050	-0.051	-0.703
	(0.039)	(0.039)	(0.611)
Selection test score (standardized)	-0.080^{***}	-0.075^{***}	-0.041
	(0.020)	(0.022)	(0.032)
Country IMO rank		-0.001	-0.001
		(0.001)	(0.001)
Competition experience		-0.011	-0.004
		(0.011)	(0.015)
Female		0.074	0.014
		(0.049)	(0.070)
Age		0.001	-0.014
		(0.018)	(0.026)
Lives in capital city		0.002	0.031
		(0.040)	(0.059)
Socioeconomic index		0.015	-0.006
		(0.017)	(0.026)
Parent has university degree		0.003	0.006
		(0.047)	(0.070)
English fluency: beginner		0.034	0.170
		(0.076)	(0.113)
English fluency: intermediate		-0.010	-0.051
		(0.043)	(0.060)
Interactions with assigned to virtual course			
× Selection test score (standardized)			-0.073^{*}
			(0.043)
× Female			0.098
			(0.099)
× Lives in capital city			-0.061
			(0.080)
× Country IMO rank			0.000
			(0.001)
× Age			0.031
			(0.035)
× Competition experience			-0.012
			(0.021)
× Socioeconomic index			0.042
			(0.035)
× Parent university degree			-0.000
			(0.095)
× English: beginner			-0.250^{*}
			(0.148)
× English: intermediate			0.098
			(0.088)
R^2	0.027	0.035	0.053
F-test: interactions = 0 (p-value)			1.20 (0.285)

Notes: This table examines predictors of not taking the final exam. The dependent variable mean is 0.419, and the sample size is 620. The F-test reported in Column 3 tests the joint null hypothesis that all interaction coefficients equal zero. The omitted category for English fluency is Advanced. Robust standard errors in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A2: Engagement and Final Exam Completion

	Did not take final exam			
	(1)	(2)	(3)	(4)
Months of engagement (course)	-0.122***		-0.116***	
	(0.013)		(0.014)	
Share of book read (control)		-0.515***		-0.520***
		(0.099)		(0.102)
Selection test score (standardized)			-0.092***	-0.038
			(0.028)	(0.030)
Observations	310	310	310	310
R^2	0.132	0.076	0.185	0.101
Dependent variable mean	0.419	0.419	0.419	0.419
Sample	treated	control	treated	control
Controls	no	no	yes	yes

Notes: This table examines the correlation between engagement (with the course or ebook, respectively) and not taking the final exam. Columns 1 and 3 use the treated group (virtual course); columns 2 and 4 use the control group (ebook only). The engagement measure for the treated group is months of active participation in the course; for the control group, it is the share of the ebook read. Columns 3 and 4 include selection test results and the standard set of controls (country IMO rank, years of competition experience, gender, age, capital-city indicator, parental education, socioeconomic index, and English fluency). Robust standard errors in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: Effect of the Virtual Course Offer on Final Exam Score — Sensitivity to Missing-Data Assumptions

	Final exam score (standardized)				
	(1) Missing = 0	(2) Missing = p25	(3) Missing = p50	(4) Missing = p75	(5) Imputed
Assigned to virtual course	0.165**	0.137**	0.113**	0.087	0.143***
	(0.073)	(0.058)	(0.054)	(0.056)	(0.053)
Strata FE	Yes	Yes	Yes	Yes	Yes
Observations	620	620	620	620	620
R^2	0.386	0.374	0.392	0.342	0.467

Notes: This table reports intention-to-treat estimates of the effect of assignment to the virtual course under alternative assumptions about missing final-exam scores. Column 1 assigns missing outcomes a value of zero. Columns 2–4 assign missing outcomes the 25th, 50th (median), and 75th percentiles of the control-group score distribution, respectively. Column 5 imputes missing scores using predicted values from a regression of observed scores on baseline covariates. All specifications include strata fixed effects. Robust standard errors in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.