Can Innovation be Taught in Schools?

Experimental Evidence from India*

Saloni Gupta

Abstract

Childhood exposure to innovation shapes who becomes an inventor, but most children grow up without it. A randomized evaluation across 80 public schools in India shows schools can provide part of this exposure. A curriculum teaching the innovation process raised participation in an innovation competition, the independently rated quality of student ideas, and skills in exploration, idea generation, fluid intelligence, and problem solving. State graduation exam scores explain at most 10 percent of skill variation, suggesting that STEM track selection based on academic scores alone may overlook innovation-relevant talent. The program developed these skills at about $40 per student annually.

JEL Codes: I21, I25, O31, O15, J24

Keywords: Innovation, education, government, India, higher-order skills, scale

*Gupta: Brown University, Department of Education (email: saloni_gupta@brown.edu). AEA RCT Registry: AEARCTR-0009696. I am extremely grateful to the Telangana Social Welfare Residential Educational Institutions Society, Telangana Tribal Welfare Residential Educational Society, Inqui-Lab Foundation, and DAI Advisory for their leadership and collaboration in this study. I especially thank Alejandro Ganimian, Eshwar Bandi, Alex Eble, Judith Scott-Clayton, Anjali Adukia, Andreas de Barros, Carolina Concha-Arriagada, Ashutosh Bhuradia, Sidra Ali, Sandra Black, Abhijeet Singh, Guthrie Gray-Lobe, and Michael Best for their feedback and support at multiple stages of this project. I have also benefited from presenting this work at academic venues including Stanford University, Brown University, The University of Chicago, Columbia University, Helsinki Graduate School of Economics, Paris School of Economics, Ashoka University, University of Zagreb, and the Advances in Field Experiments and Association for Education Finance and Policy conferences. This project was financially supported by the Weiss Fund at the University of Chicago, the NAEd/Spencer Dissertation Fellowship, and multiple small research grants from Teachers College, Columbia University. This study received IRB approval from Teachers College, Columbia University (Protocols #22-352 and #22-310), and Stanford University (Protocol #79854). AI software (Claude, Anthropic; Codex, OpenAI) was used to assist with drafting, editing, and code development during manuscript preparation; the author is solely responsible for all content.

I. Introduction

Long-run economic growth depends on the production of new ideas (Romer 1990). Participation in this process is unequal: the probability of becoming an inventor is shaped in part by childhood exposure to innovation and varies sharply by family income, gender, and geography (Bell et al. 2019; Hoisl, Kongsted, and Mariani 2023).[1] Such childhood exposure may make inventive careers visible, reveal how ideas are developed, and create opportunities to engage in that process. Extending this exposure to more children is also relevant in low- and middle-income countries, where formal research workforces are much smaller. India reported 259 full-time-equivalent R&D researchers per million people in 2020, compared with 490 in lower-middle-income economies and 4,149 in high-income economies (World Bank 2024b, 2026b).[2] Schools cannot reproduce the relationships and environments that ordinarily provide exposure to innovation, but their broad reach may allow them to offer one component at scale: repeated, guided practice in developing and testing ideas.

This paper asks whether such instruction builds students' skills in generating and developing new ideas.[3] I study this question in a randomized controlled trial across 80 government residential schools in Telangana, India. The schools serve students from Scheduled Castes and Scheduled Tribes—historically marginalized caste and tribal communities—whose fathers predominantly work as farmers or daily-wage laborers. Of the 80 schools, 40 received the program and 40 served as controls. The study follows the 5,642 students enrolled in grade 8 for three academic years, through the high-stakes grade 10 state graduation examinations. The program, implemented by the Government of Telangana in partnership with the Inqui-Lab Foundation, comprised four instructional cycles over two years. In each cycle, students identified problems in their communities, interviewed the people affected, generated and compared possible solutions, and, finally, built

[1] Bell et al. (2019) show that children from top-1-percent families are ten times as likely to become inventors as those from below-median families. Even among children in the top 5 percent of third-grade mathematics scores, those from high-income families are more than twice as likely to invent. Children disproportionately patent in technology classes represented among inventors in their families and neighborhoods; girls are more likely to enter classes in which they encountered female inventors. Hoisl, Kongsted, and Mariani (2023) document inequality within inventor families: about half the inventor-parent advantage for first-born daughters disappears when their younger sibling is a boy rather than a girl, consistent with gendered transmission of occupation-specific support. Agarwal and Gaulé (2020) document a related institutional channel: conditional on International Mathematical Olympiad performance, participants from lower-income countries are less likely to enter mathematics PhD programs and produce less research, with weaker domestic training and research capacity accounting for part of the gap.

[2] India's resident patent applications number 19 per million people, compared with roughly 1,800 in Japan and 3,600 in Korea (World Bank 2023). Within its 3i framework, the *World Development Report 2024* identifies skilled labor as essential to technology infusion and innovation, and broad access to education as a means of expanding the talent pool. It cites Finland's comprehensive-school reform as broadening that pool and potentially weakening the relationship between parental income and later inventorship (World Bank 2024b, pp. 63–64, 189–194).

[3] The Oslo Manual defines an innovation as "a new or improved product or process (or combination thereof) that differs significantly from the unit's previous products or processes and that has been made available to potential users (product) or brought into use by the unit (process)" (OECD and Eurostat 2018). Student ideas and prototypes satisfy the first part of this definition, but not the implementation requirement. The study emphasizes novelty and usefulness at the idea and prototype stages; *innovation* refers here to this early production rather than patenting or commercialization.

and revised prototypes using contemporary technology and material kits. Two trained student leaders in each classroom led the sessions using structured workbooks, and teachers had a supporting role. Students in all 80 schools could submit ideas to the Innovation Challenge, which operated under the same announcements, deadlines, and prizes; shortlisted teams from either arm pitched their ideas to an expert jury. The design therefore varied access to guided practice while holding fixed the formal opportunity to enter the competition.

Evaluating school-based instruction in innovation raises three challenges. First, innovation is difficult to measure at this stage of schooling: patents and commercialization, the conventional measures of innovation, are not observable, and there is no established standardized and scalable measure of the ideas students produce. Second, students with weaker foundational academic skills may benefit less from instruction in higher-order innovation skills. Dynamic complementarity in skill formation implies that skills acquired earlier can raise the productivity of later investments (Cunha and Heckman 2007).[4] This possibility is especially relevant in India, where 54 percent of rural grade 8 students cannot solve a division problem set at the grade 3 to 4 level (ASER Centre 2025). Third, the program's 51 contact hours may displace time for core academic learning, particularly because instruction takes place during grades 8 and 9, immediately before the grade 10 board examinations.

To measure the novelty and usefulness of the ideas students produce, the study constructs Innovation-S by translating the Oslo Manual's definition of innovation into seven five-point items (OECD and Eurostat 2018). The items draw on interviews with 14 practicing innovators and a survey of 44 adults building products they regard as innovative. In the study's second year, teams from treatment and control schools submitted 1,225 ideas to the Innovation Challenge using a common format: a drawing and an audio-recorded pitch. All pitches were transcribed, and blinded evaluators scored the transcripts using Innovation-S. At the finals, professional jurors assessed the pitches using the same rubric and allocated funding among the teams. The within-juror correlation between jurors' rubric scores and their funding allocations was 0.77.

Innovation is a multistage process that also requires component skills: exploring alternatives, generating solutions, diagnosing design problems, and building and testing prototypes. To measure these skills, each student completed four assessments: an incentivized search task for exploration (March 1991; Ederer and Manso 2013), PISA creative-thinking tasks for idea generation (OECD 2024), the ICAR matrix battery for fluid intelligence (Condon and Revelle 2014), and Sudoku puzzles for well-defined problem solving (Newell and Simon 1972). Responses to the creative-thinking tasks are open-ended text in English or Telugu, scored against task-specific rubrics by a large language model. A research assistant blind to treatment status verified

[4]Pritchett and Beatty (2015) show how curricula that advance faster than students' mastery can generate little learning in developing-country settings. In education practice, a common interpretation of Bloom's taxonomy treats higher-order objectives as building on lower-order ones (Bloom et al. 1956; Anderson and Krathwohl 2001).

all assigned scores and confirmed more than 98 percent without changes. Because the assessments use tasks distinct from those practiced in the curriculum, gains on them provide evidence that the program developed the underlying skills.

I report six main results. First, the academic scores used to screen students into STEM education are weak predictors of innovation-relevant skills. Many senior-secondary schools use grade 10 board scores to admit students to the science and computer-science streams of grades 11 and 12, the principal school pathways into STEM degrees (CBSE 2010; Delhi DoE 2026). Among control students, the baseline academic score and the grade 10 board score each explain no more than 12 percent of the variation in idea generation, problem solving, fluid intelligence, or exploration (Figure 1). The profiles are relatively flat: higher academic scores are associated with only modest increases in these skills. A screen limited to the top academic quintile would miss 60 to 80 percent of the students who rank in the top quintile on these skills.

Second, structured instruction changed both who produced innovative work and the quality of what they produced. Treatment raised participation in the Innovation Challenge by 8.0 percentage points, from a control mean of 73.3 percent. Blinded evaluators rated the ideas of treatment-school teams higher in novelty and usefulness, by 0.187 and 0.146 standard deviations. Among submitted ideas, the effect is present in each of the two evaluators' scores separately. Funding decisions at the competition's finals are consistent with the evaluators' ratings: eight of the ten highest-funded teams were from treatment schools (one-sided randomization-inference $p = 0.075$). The program's effects on participation and idea quality show that schools can broaden early exposure to the process of idea generation and development.

Third, treatment improved four higher-order skills. The exploration index rose by 0.094 index points, idea generation by 0.314 standard deviations, fluid intelligence by 0.111, and problem solving by 0.142. The gains show transfer beyond the curriculum's projects. Students practiced exploring alternatives, generating solutions, and building prototypes for community problems, and improved on standardized assessments of the underlying skills.

Fourth, treatment produced no detectable change in academic performance. At endline, the effects on mathematics and science are -0.007 and 0.027 standard deviations and statistically indistinguishable from zero. On the grade 10 state graduation examinations a year later, administrative records cover nearly the entire cohort. Estimated effects on language scores are positive and significant at the 10 percent level before adjustment, but no subject-level effect remains statistically significant after adjustment for multiple testing. Students reported allocating a greater percentage of their weekly time to the Innovation Challenge and sports and a smaller percentage to other extracurricular activities, including dance, singing, drawing, and reading. The percentages devoted to school periods and homework showed no detectable reduction.

Fifth, students with weaker baseline academic performance made substantial gains in participation and innovation-relevant skills. Participation rose by 9.7 to 12.0 percentage points in the bottom three academic quintiles, against an insignificant 3.3 points in the top. Idea generation improved in every quintile, with the largest gain—0.64 standard deviations—in the top quintile; exploration gains were similar across quintiles.

Sixth, the direction of submitted ideas differs by gender. Girls' teams direct a larger share toward people and care (public welfare, health, and households), while boys' teams direct a larger share toward the environment and transport.[5] Because all schools are single-gender and students in girls' schools scored 0.36 standard deviations higher on the baseline academic assessment, these comparisons are descriptive and are not experimentally identified effects of gender.

This study contributes to two bodies of literature. The first concerns the range of skills schools can develop and the instructional approaches through which they can do so. Foundational literacy and numeracy receive strong emphasis in education policy and spending (Evans and Hares 2021; World Bank 2024a; CGD 2025); India's National Education Policy names them the highest priority of the education system (Government of India 2020, para. 2.2). Yet economically valuable human capital extends beyond these foundations: noncognitive, social, and higher-order skills also shape educational attainment, employment, and earnings (Heckman, Stixrud, and Urzua 2006; Heckman and Kautz 2012; Deming 2017, 2022). Schools can develop these broader skills. Interventions fostering grit (Alan, Boneva, and Ertac 2019), curiosity (Alan and Mumcu 2024), and cognitive endurance (Brown, Kaur, Kingdon, and Schofield 2025) improve academic performance. The benefits can persist beyond instruction, with higher educational attainment ten years after negotiation training (Ashraf, Bau, Low, and McGinn 2020; Ashraf, Bau, Low, and Shan 2025) and better business performance nine years after entrepreneurship training (Chioda, Gertler, Contreras-Loya, and Carney 2026). Beyond interventions targeting particular skills, research shows that changes in pedagogical practice and students' roles in instruction can improve educational outcomes. Inquiry- and problem-based approaches improve learning by engaging students in questioning, investigation, and problem solving (Bando, Näslund-Hadley, and Gertler 2019; Nourani, Ashraf, and Banerjee 2025), while assigning students teaching responsibilities can improve their behavior and educational progression (Alan and Kubilay 2025).

Together, these studies document that schools can develop skills beyond conventional academic objectives. Innovation is central to economic growth (Romer 1990; Jones 2022). Building on this research, this paper provides experimental evidence that schools can develop human capital for innovation. Conventional

[5]The tilt of girls' teams toward people-focused problems is not specific to this sample. Among 327,599 submissions to the 2025 round of the national School Innovation Marathon, conducted that year as the Viksit Bharat Buildathon, 58.7 percent of teams submitting in women-and-child-development themes are girl-majority, against 50.0 percent of teams overall (Inqui-Lab Foundation 2026).

academic tasks typically specify the problem and the criteria for a correct answer, whereas innovation involves ill-defined problems for which neither the problem formulation nor the solution is given. This paper shows that a standardized school curriculum can also support open-ended production, in which students must define the problems themselves and develop and revise their own solutions without a predetermined answer. It shows that government schools can teach a structured process for innovation without specialist teachers or dedicated laboratories.

The second contribution is to the literature on the allocation of talent to innovation. The study provides evidence on the relationship between academic achievement, innovation-relevant strengths, and gains from instruction. Family background and exposure to innovation shape who becomes an inventor (Bell et al. 2019; Aghion, Akcigit, Hyytinen, and Toivanen 2017; Toivanen and Väänänen 2016), and barriers that prevent talented individuals from entering innovation can impose substantial growth costs (Brouillette 2025). Research on talent identification shows that exceptional performance in science competitions can identify students with strong prospects for subsequent scientific achievement (Agarwal, Gaulé, and Jiang 2025). This study complements that evidence by showing that ordinary academic achievement provides an incomplete account of innovation-relevant strengths across a school cohort. Students with weaker academic performance also develop innovation-relevant skills through instruction. These findings have implications for the allocation of opportunities to learn and practice innovation: screening on academic achievement alone may overlook students with relevant strengths who can benefit from such opportunities.

The study also provides a framework for assessing student-produced ideas and individual innovation-relevant skills, combining independent ratings with standardized, performance-based assessments. Early innovation has no established measure, and measures of noncognitive skills developed in high-income countries have shown low reliability in low-income settings (Laajaj and Macours 2021). Within the individual assessments, the measurement of idea generation complements the divergent-thinking tasks commonly used in creativity research, such as the Alternate Uses Task (Guilford 1967). Open-ended tasks adapted from the PISA creative-thinking assessment, first administered in 2022 (OECD 2024), are scored at scale using a large language model. This approach makes open-ended responses usable in large-sample evaluations while retaining task-specific assessment criteria.

The findings inform innovation and education policy, particularly the design of school innovation programs that governments are already promoting. India has placed innovation laboratories in schools across the country and conducts a nationwide student ideas competition (Atal Innovation Mission 2024; Press Information Bureau 2024). The governments of Delhi and Andhra Pradesh have introduced an entrepreneurship-mindset curriculum in their government schools (SCERT Delhi 2021; Amaral et al. 2025). Rwanda requires all

secondary students to take an entrepreneurship course (Blimpo and Pugatch 2021), and Kenya's grade 9 curriculum includes a project in which students identify a local problem and make an item to solve it (Kenya Institute of Curriculum Development 2024). Governments also spend billions of dollars each year on research and development and on startup ecosystems (World Bank 2026a; World Intellectual Property Organization 2024). The effect of this spending on innovation depends on the supply of people with the human capital to develop new ideas (Bloom, Van Reenen, and Williams 2019; Van Reenen 2021), and this study shows that schools can start to build that human capital.

The evaluation supports these policy implications with evidence on delivery and measurement at an intermediate implementation scale (Al-Ubaydli, List, and Suskind 2017; List 2024; Muralidharan and Singh 2025). The program was delivered within the ordinary operations of 40 government schools, and the Inqui-Lab Foundation monitored implementation. The assessment battery was also administered at scale, on tablets to 5,215 students at baseline and 4,388 at endline across the 80 schools. With support from the Agency Fund, the battery is now being extended to participants in the national School Innovation Marathon, whose 2025 round drew 327,599 submissions (Inqui-Lab Foundation 2026).

The remainder of the paper proceeds as follows. Sections II–V describe the setting and intervention, experimental design, outcomes, and empirical strategy. Section VI presents the results, Section VII discusses cost-effectiveness and scalability, and Section VIII concludes.

II. Context and Intervention

A. Setting

The experiment was conducted in Telangana's government residential schools for students from Scheduled Castes (SCs) and Scheduled Tribes (STs), two constitutionally recognized groups that have experienced persistent historical disadvantage. Scheduled Castes (also known as Dalits) were historically subjected to untouchability and exclusion under the caste system, while Scheduled Tribes (also known as Adivasis) are indigenous communities that have long faced geographic isolation and social and economic marginalization. The schools belong to two parallel state networks, Social Welfare Residential Schools, which predominantly serve Scheduled Caste students, and Tribal Welfare Residential Schools, which predominantly serve Scheduled Tribe students; the two societies that administer them operate more than 430 single-gender institutions in total, from primary schools through undergraduate colleges (TGSWREIS 2026; TGTWREIS 2026).

The households these schools serve depend largely on manual and informal work. Caste continues to shape occupational sorting, social networks, and access to credit, limiting economic opportunities across generations (Munshi 2019); caste-based wealth gaps are large and persistent (Bharti 2018), and intergenerational mobility

is low, despite the gains from affirmative action (Asher, Novosad, and Rafkin 2024). Exposure to technical and inventive work, which children may otherwise acquire through family occupations or social networks, is less likely in these settings.

Students in these residential schools are, however, academically selected from disadvantaged populations. Admission is selective, determined through a statewide entrance examination: in the 2026 cycle, about 173,000 students took the examination and about 48,000 were selected for admission across the state's residential school societies (Government of Telangana 2026; Deccan Chronicle 2026; Sihag 2026). Once enrolled, students in both networks perform well on high-stakes state assessments. In the 2022 grade 10 board examinations, pass rates were 98.1 percent in the social-welfare schools and 96.0 percent in the tribal-welfare schools, against a statewide rate of about 90 percent (TSWREIS 2022; TGTWREIS 2024); these figures reflect the selectivity of admissions as much as the schools' own contribution. These students are well prepared by conventional academic measures.

B. The Innovation Curriculum

Treatment schools received Think & Make, a two-year structured curriculum that teaches the process of innovation directly. The Inqui-Lab Foundation developed the curriculum and has delivered it in these welfare schools since 2017, in partnership with the Government of Telangana; the foundation now operates in four states. The curriculum comprises four blocks over the two study years (grades 8 and 9), two per year: Planet and People in Year 1, and Design Evolutions and Smarter Everyday in Year 2. Each block is a sequence of seven to nine weekly 90-minute sessions, 51 contact hours in total. Students work through the sessions in teams of four; teams of three or five were rare. Appendix Table B.1 summarizes the four-block sequence, required student work, and classroom delivery model.

Each block follows the same five-stage cycle: find, learn, ideate, make and test, and present. Teams observe their school, homes, and public spaces and interview the people affected to select a problem (find); study how existing products and solutions evolved (learn); generate multiple solutions under time and material constraints and shortlist them against criteria (ideate); build a prototype from kit materials, gather feedback from at least two users, and revise (make and test); and present how they developed and tested their idea to their class and school (present). Prototypes are built and rebuilt across sessions; failure is expected, and teams debug, redesign, and iterate.

Each stage of the cycle both advances the team toward a working prototype and provides practice in a distinct innovation-relevant skill. Finding problems through observation and interviews practices perspective-taking; studying how designs evolve practices pattern recognition; generating and shortlisting solutions

practices divergent thinking and evaluative judgment; building and revising prototypes practices structured problem solving and persistence; presenting and absorbing feedback practices communication. Repeating these demands over four cycles is the mechanism by which the curriculum is designed to build innovation-relevant skills. Inqui-Lab's session guides map each stage to the skills it is meant to build; the research team then pairs each skill with an established assessment from the relevant discipline. Section IV describes this mapping further.

The curriculum becomes progressively more open-ended as students repeat the cycle. In Year 1, teams use guided story cards with step-by-step prompts and structured evaluation rubrics; in Year 2, they progress to semi-open timed design challenges and then to fully independent problem discovery, in which teams define both the problem and the solution themselves. The final block applies the same problem-solving cycle to a new medium, sensor-based prototyping, with students wiring sound, light, heat, and distance sensors to outputs such as LEDs, motors, and buzzers. Each classroom received an innovation kit, valued at Rs 15,000 (about $180), containing circuits, sensors, recyclable materials, and beginner-friendly prototyping tools. Appendix Figure B.1 illustrates this progressive release of autonomy using tasks from the student workbooks.

The program used a student-led active-learning model. Sessions used structured student workbooks containing step-by-step prompts, reflection sheets, and peer-feedback rubrics that allowed teams to work semi-autonomously. Two student leaders per classroom, nominated by teachers and given 20 hours of training, facilitated the sessions, managed time and materials, and were instructed to answer only clarifying questions.

The choice of student-run delivery reflected the constraints of the government school system. Teachers in these schools face high student-teacher ratios and substantial administrative duties. The program therefore assigned them a limited supporting role. A teacher was expected to be present during sessions but did not deliver the curriculum. The teacher supported the student leaders as needed, mediated when problems arose, advised teams, and helped organize the end-of-block showcases. Neither the student leaders nor the teacher supplied problems or solutions; defining the problem and producing the solution remained the teams' own work. Sessions were held during the after-school study period or in place of other non-academic periods. The curriculum did not directly teach or test mathematics, science, or other examination-subject content. Its designers consulted India's national science textbooks to keep activities within the range of material students had already encountered.

Schools completed 98.9 percent of scheduled sessions over the two program years. The schedule comprised 18 sessions per treatment school in Year 1 and 16 sessions per treatment school in Year 2. Across the 40 treatment schools, 1,345 of the 1,360 scheduled sessions were delivered. In each program year, 35 schools completed every session scheduled across both structured units. Completion status was compiled

TABLE 1—IMPLEMENTATION AND MONITORING OF THINK & MAKE

	Year 1		Year 2		Tracked total
	Planet	People	Design Evolutions	Smarter Everyday	
Sessions scheduled per school	9	9	9	7	34
Scheduled school-session records (sessions per school × 40 schools)	360	360	360	280	1,360
Completed school-session records	359	354	357	275	1,345
Session completion rate	99.7%	98.3%	99.2%	98.2%	98.9%
Field-team observer present, *N* (% scheduled sessions)	215 (59.7%)	183 (50.8%)	146 (40.6%)	83 (29.6%)	627 (46.1%)
Schools completing every scheduled session	39/40	35/40	37/40	38/40	35/40 in each year

Notes: The table reports implementation across the 40 treatment schools. Each structured unit scheduled seven or nine sessions per school, producing the school-session totals shown in the table. Completion status was compiled from progress reported by the student leaders and observers' reviews of student manuals during school visits. Seven field-team observers covered the 40 schools over the two program years. The observer row reports sessions at which a field-team observer was present. In each program year, 35 schools completed every session scheduled across both structured units.

from progress reported by the student leaders and audits of student manuals during school visits.

The program paired student-run delivery with regular external monitoring. Seven field-team observers covered the 40 treatment schools over the two years. They observed 627 sessions, or 46 percent of the scheduled curriculum, and used a standardized observation rubric to record classroom implementation.

Classroom observations show student leaders managing sessions and most teams completing the planned work. An observed section enrolled approximately 40 students, who formed nine to ten teams of four depending on the day's attendance; observers recorded most or all teams completing the session's activities in 83 percent of observed sessions. In the Planet unit, for example, the problems teams selected and carried through the cycle were drawn from the students' surroundings, most often the waste of water and food, and local pollution. The two student leaders led the session: they opened it by reviewing the previous session's work, announced the day's goals, managed the time, circulated among the teams to encourage participation and answer clarifying questions, and closed the session. Teachers were present in approximately two-thirds of the classroom observations for which their presence was recorded; for context, nationally representative evidence from rural Indian public schools documents a teacher absence rate of 23.6 percent during unannounced visits (Muralidharan, Das, Holla, and Mohpal 2017). Table 1 reports completion and observation by curriculum unit.

The program cost approximately $40 per student per year; Section VII examines cost-effectiveness and scalability.

III. Experimental Design and Data

A. School Selection and Randomization

The study sample comprises 80 schools, drawn from the 320 secondary schools serving grade 8 that the two societies operate. A school was eligible if no other externally run program was active when the study began, so that the treatment contrast would not be confounded by a concurrent intervention, and if it was located in a district that Inqui-Lab had shortlisted as operationally feasible for the in-person support its field team provides. The 80 schools span 14 districts and every institutional type in the system: social-welfare and tribal-welfare management, girls-only and boys-only composition, and rural and urban locations. Figure 2 summarizes the research design and timeline.

Treatment was assigned at the school level, before the baseline assessment, within eight strata formed by the intersection of the three characteristics above: management, gender composition, and location. Within each stratum, half of the schools were randomly assigned to treatment and half to control, yielding 40 treatment and 40 control schools that are balanced on school characteristics across arms (Appendix Table A.2).[6]

B. The Innovation Challenge and the Control Condition

The Innovation Challenge was a team competition held at the end of each school year in all 80 study schools. It was identical in the two arms: the two management societies issued the same circulars, and the same deadlines and prizes applied. Students in both arms formed teams of four to identify a problem and propose a solution. The announcements stated that selected teams would be invited to a Shark Tank-style finals in the state capital, where they could win funding (Section IV); for these residential-school students, most in rural areas, the trip to the capital was part of the opportunity. The format was familiar: *Shark Tank India* had aired three national seasons before the study's Year 2 finals,[7] and Indian business press describes the show as encouraging young people to pursue entrepreneurship (Khan 2024).

The control condition paired this competition with administrative encouragement. In control schools, headmasters and teachers were informed through WhatsApp that surveyors would visit at the end of each school year and that every team would be asked to submit an idea, and the circulars and WhatsApp messages announced the criteria on which ideas would be judged, so control teams were informed of the assessment criteria. Control schools received no instruction in how to develop an idea: teams prepared their submissions

[6]The study is registered at the AEA RCT Registry (AEARCTR-0009696; Gupta 2025); Appendix G lists deviations from the registered pre-analysis plan.

[7]Shark Tank is a televised pitch format in which entrepreneurs present business ideas to a panel of investors who may fund the venture; the Indian version debuted in 2021 and has aired five national seasons (SonyLIV 2026; Jha 2025).

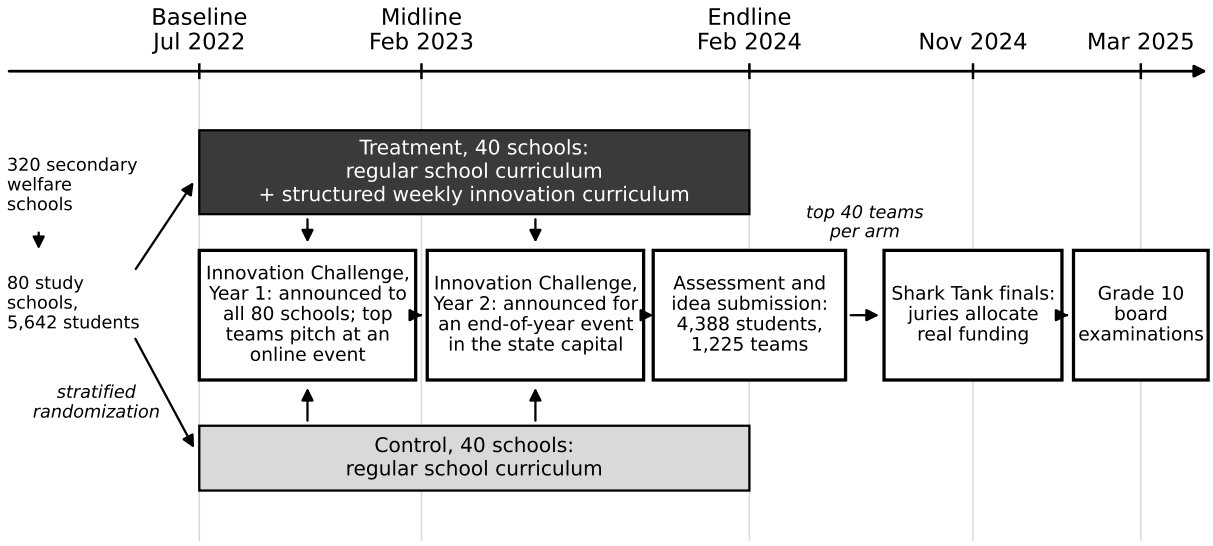


FIGURE 2. RESEARCH DESIGN AND TIMELINE

Notes: The figure shows the study timeline for the 80 study schools, drawn from 320 secondary welfare schools and randomized within strata into 40 treatment and 40 control schools. The Innovation Challenge was announced to all 80 schools at the start of each school year; top-rated teams from both arms pitched at an online event at the end of Year 1. Idea submissions are measured for the Year 2 cycle and were collected during the endline school visit, in which 4,388 students completed the individual endline assessment and 1,225 teams submitted ideas. The top 40 teams from each arm were invited to the Shark Tank finals, where juries allocated funding among the ideas they heard. Grade 10 board examinations followed in March 2025. The horizontal axis is not drawn to scale.

with the support available from their teachers and headmasters. Control schools did not receive the innovation kits during the program years; once the finalists were announced, the 40 control finalist teams received the same kits. Finalist teams in both arms had two weeks to prepare, the length of the curriculum's prototype module. School innovation competitions are an established policy instrument in India: Telangana's School Innovation Challenge, run with UNICEF since 2020, and the national School Innovation Marathon, launched in 2024 and followed in 2025 by the Viksit Bharat Buildathon, invite students to present solutions to practical problems (Telangana State Innovation Cell 2026; Dejuan and Sanz Navarro 2023; Press Information Bureau 2024, 2025). The control condition therefore approximates the prevailing low-cost model for encouraging innovation in schools: an announced competition, official encouragement, and reliance on existing school staff.[8]

The contrast between arms measures what the structured curriculum adds over announcement and encouragement: whether it changed take-up of the opportunity and the quality of submitted ideas.

[8]The study team did not tell control schools about the curriculum, but headmasters may have learned through the societies' administrative networks that Inqui-Lab was delivering a program in other schools. If that awareness, or the announced competition itself, raised effort in control schools, the treatment contrast understates the curriculum's effect relative to a no-competition counterfactual.

C. Data Collection

The paper's outcomes are measured across three waves of individual tablet assessments. The baseline, administered before the program began, measured academic achievement, fluid intelligence, emotional intelligence, risk preferences, Big Five personality, and exploration behavior on the incentivized search task. The midline, administered in February 2023 near the end of Year 1, repeated the baseline instruments. The endline, at the end of Year 2, administered new variations of the same instruments and added creativity and problem solving, together with the Innovation Challenge. The Challenge was conducted in both program years, but only the Year 2 submissions are scored: the Year 1 submissions proved unusable at similar rates in both arms,[9] and the protocol was redesigned for Year 2 to capture the students' own audio-recorded pitches. Appendix Table A.1 reports the number of students assessed in each wave.

Tablet administration made it possible to measure interactive search and problem solving alongside academic performance. The assessments included incentivized tasks such as the search task and timed Sudoku rounds. Behavioral measures of this kind, from grit to negotiation, have been refined in field experiments and used to document meaningful impacts (Alan, Boneva, and Ertac 2019; Ashraf, Bau, Low, and McGinn 2020); the next step for research on higher-order skills is to take these assessment instruments to national scale, and tablet administration makes that step feasible.

Tablets also improved data integrity in several ways. Module order was randomized across students, so adjacent students encountered different content, and each item appeared on its own screen, which limits copying even among students seated side by side (Singh 2024). Every item required a response before the assessment advanced, which held item-level missingness low: a disengaged student could guess and proceed, and could stop the assessment by returning the tablet, but could not leave items blank. Responses were recorded directly in the survey application and transmitted to the server, avoiding the read errors common to optically scanned paper answer sheets. Because the instruments were digital, they could be piloted and revised rapidly in out-of-sample schools before each wave: an item that underperformed was revised or dropped within two days, faster than paper-based instruments typically allow.[10]

[9] In Year 1, surveyors typed each team's pitch in English while the students spoke, translating in real time from Telugu, the native language of both the students and the surveyors; records produced this way could not be scored reliably. The Year 1 cycle nonetheless concluded as announced: the typed records were sufficient to shortlist top teams, and shortlisted teams from both arms pitched at an online finals event held by video conference. Because the clarity of a team's record may itself have affected shortlisting, no idea-quality outcome is constructed from the Year 1 cycle.

[10] The assessments were administered using the Qualtrics Offline App, so administration did not require an internet connection. Enumerators worked in teams of five: at each school they registered students by name and unique identifier, distributed pre-charged tablets loaded with the battery in English and Telugu, and supervised the session in a designated classroom. Enumerators received three days of training and followed a standardized script, so instructions were delivered consistently across schools, and they did not assist students with the content of their answers; after each visit the field coordinator checked the synchronized responses for completeness.

The endline visit also collected the team submissions to the Innovation Challenge: 1,225 teams submitted ideas on the day of the individual assessment. The individual assessment yields the student-level outcomes; the team submissions yield the participation and idea-quality outcomes, defined in Section IV. Appendix Table A.1, Panel B, reports the division of the 4,388 endline-assessed students into submitters and non-submitters and the construction of the team analysis sample. Students were assigned to teams of four at random after baseline; most teams had reorganized by the Year 2 competition, and the share of assigned teams with unchanged membership was similar across arms (22.9 percent treatment, 21.9 percent control, $p = 0.71$), so team composition at endline is largely self-selected in both arms.

D. Sample, Balance, and Attrition

The study cohort comprises the 5,642 students enrolled in grade 8 across the 80 schools at baseline: 2,835 in treatment schools and 2,807 in control schools, approximately 70 per school. The baseline assessment was completed by 5,215 of these students, or 92.4 percent (2,618 in treatment schools and 2,597 in control schools).

Treatment and control students were balanced on observed baseline characteristics in the sample of 5,215 baseline-assessed students (Table 2). Across the composition rows and ten student covariates, no treatment-control difference is significant at conventional levels, and joint tests do not reject balance ($p = 0.48$ for student characteristics, $p = 0.97$ for baseline assessment measures).

Differential attrition would threaten the experimental comparisons. I test for it on two rosters: the baseline roster of every student assessed before the program began, and the grade 10 board examination records, which list every student still enrolled in the school system three academic years later. Baseline-assessed students are linked to a board examination record at statistically indistinguishable rates in the two arms: 77.7 percent of control and 76.8 percent of treatment students ($p = 0.68$).

Attendance at each wave, reported in Appendix Table A.1, does not differ significantly across arms: 98 percent of control students attended midline and 76 percent attended endline, with treatment rates 4.1 percentage points lower and 4.1 points higher respectively, neither difference significant ($p = 0.18$ and 0.12).

The attrition checks detect no differential selection on observed baseline characteristics. No baseline covariate predicts endline observation differently by arm, individually or jointly ($p = 0.50$). Students not observed at endline do not differ significantly from those observed in their externally graded board scores: 0.04 standard deviations higher in mathematics, 0.03 lower in science. Students assessed at endline are also balanced across arms on all nine baseline covariates (joint $p = 0.80$). Appendix Table A.3 reports all four tests.

TABLE 2—EXPERIMENTAL SAMPLE AND BASELINE BALANCE

	Control mean	Treatment mean	Treatment− Control	*p*-value	*N*
Design					
Randomized schools	40	40			80
Students assessed at baseline	2,597	2,618			5,215
Panel A. Student characteristics					
Female students (%)	54.4	51.6	-2.8	0.807	5,215
Tribal-welfare school (%)	48.0	48.2	0.2	0.984	5,215
Rural school (%)	62.1	63.0	0.9	0.937	5,215
Age (years)	13.336	13.285	-0.061	0.169	5,215
Household SES index (SD)	0.054	-0.048	-0.114	0.212	3,566
Household computer literacy (0–1)	0.381	0.401	0.025	0.396	4,029
Joint test *p*-value				0.483	3,566
Panel B. Baseline assessment measures					
Mathematics score (%)	41.6	41.0	-0.4	0.837	5,215
Science score (%)	52.4	52.4	0.1	0.884	5,215
Fluid intelligence (% correct)	57.7	57.9	0.2	0.900	5,215
Emotional intelligence (% correct)	52.0	51.4	-0.5	0.552	5,215
Exploration index (SD)	0.000	-0.021	-0.020	0.591	5,215
Big Five (mean item response)	3.18	3.17	-0.01	0.694	5,215
Risk preference (% of endowment)	63.0	63.4	0.4	0.689	5,210
Joint test *p*-value				0.974	5,210

Notes: This table reports baseline summary statistics and balance. The student sample is students who completed the baseline assessment. Each row reports means by treatment arm; the treatment-control difference is the coefficient on treatment assignment from a regression of the covariate on treatment and randomization strata, with standard errors clustered at the school level; *N* is the number of students with nonmissing values for that covariate. The female, tribal-welfare, and rural rows are the stratification variables and report raw differences without strata controls. All schools are single-gender residential schools run by one of two government societies: social-welfare schools serve predominantly Scheduled Caste students, and tribal-welfare schools serve predominantly Scheduled Tribe students. School characteristics and their balance are reported in Appendix Table A.2. Joint test *p*-values are from Wald tests of all panel covariates in a regression of treatment assignment on the covariates and randomization strata. The household SES index is the first principal component of eight household asset indicators (air conditioner, car, refrigerator, laptop, scooter, washing machine, and father's and mother's mobile phones), standardized; non-imputed values only. Household computer literacy equals one if a member of the student's household knows how to use a computer; the SES index and household computer literacy are measured in the midline household modules. Mathematics, science, fluid intelligence, and emotional intelligence are percent correct on the baseline instruments. The exploration index mirrors the endline exploration measure: the equal-weighted mean of three search behaviors from the round-level baseline exploration game (initiation, discovery, and convergence), each standardized to the control group, with convergence reversed so that higher values indicate more and better-directed search. Because the components are standardized to the control group, the control mean of the index is zero by construction. The Big Five row is the mean item response across the five trait scales, on the 1–5 item scale with neuroticism reversed and the traits equally weighted. Risk preference is the average share of the endowment bet across nine incentivized rounds, following the investment game of Gneezy and Potters (1997).

IV. Outcomes and Measurement

This section describes the outcome measures in four groups, in the order the results report them. The Innovation Challenge provides measures of participation in the competition and the quality of submitted ideas. An incentivized search task measures exploration, one of the four component skills, by recording

search behavior directly rather than through an item-based test. The remaining skill tasks measure the other component skills the curriculum practices, each with a separate instrument. Academic outcomes are measured in the endline assessment and the grade 10 board examinations. Throughout, item-based scores are estimated with a two-parameter item-response model and standardized to the control group.

A. The Innovation Challenge and the Innovation-S Scale

Submitting teams were asked to list all team members, including teammates who were absent on the day of submission. I use those rosters to construct participation: among students who completed the individual endline assessment, administered on the day submissions were collected, a student is coded as participating if named on a submitted team roster.

Each submission consisted of a drawing of the idea and an audio recording describing the problem the team wanted to solve, the proposed solution, the intended users, and why the solution improved on existing alternatives; teams were not permitted to submit prototypes, so submissions took the same form in both arms. The evaluators score the idea itself, and all scored submissions are from the Year 2 cycle.[11] Appendix Figure C.1 shows a submitted drawing. Teams selected one of nine sectors on the submission form. I classify intended users from a separate short written response using keyword rules described in Appendix Figure A.3.

To score the submissions, I developed a new instrument for this study, the Innovation-S scale. The scale operationalizes the Oslo Manual's definition of innovation, a new or improved product made available to its users (OECD and Eurostat 2018), as a rubric that can be applied to a school team's idea. Its items were developed from interviews with 14 adult innovators, including a Nobel laureate and three patent holders, and a survey of 44 adults involved in developing products they regard as innovative, including founders, product managers, and executives.

The main scoring protocol used two trained evaluators to score submissions independently on four measures. Novelty and value-add average four and three five-point items, respectively (Appendix Table C.1). Value-add measures the usefulness of an idea to its intended users. Evaluators also rated feasibility on a five-point scale and stated their willingness to fund each idea, a hypothetical amount between Rs 0 and Rs 1,000. Evaluators read the submissions sorted by sector and were not informed of a team's name, school, or treatment status.[12] The rubric measures the same dimensions teams were told to aim for: the guidelines circulated to teams instructed them to emphasize the novelty of the solution, its benefit to intended users, and its accessibility "considering cost and effort," and the announced judging criterion was "how beneficial the

[11] Year-1 submissions could not be scored reliably (Section III.C); the Year 2 protocol recorded each team's pitch, with the recordings transcribed afterward and the transcripts provided to the evaluators.

[12] The two evaluators were recruited locally by the data collection firm rather than by the research team, and I had no direct contact with them.

innovative idea is for its intended users." The ratings therefore measure idea quality on criteria announced identically to both arms, assigned by evaluators blind to treatment status.

These measures separate participation in the Innovation Challenge from the rated content of submitted ideas. The main results report both margins: whether treatment increased participation, and whether it increased evaluator-rated idea quality after accounting for the fact that treatment also affects submission. Appendix C gives additional detail on the scoring protocol and rater validation and shows submitted ideas across the quality distribution (Figure C.2); the conditional submitted-team estimates appear in Appendix A.

The Innovation Challenge concluded with a Shark Tank-style finals event in Hyderabad, where professional jurors disbursed funding among the ideas they heard. The 40 highest-rated teams from each arm were invited, with travel costs covered; two invited control teams did not attend. The number of invitations was fixed at forty teams per arm; funding comparisons concern these selected finalists. Finalists in both arms received the same materials and preparation time before the event. At the event, each team pitched its idea live to a panel of professional jurors, typically two per team. Each juror could allocate up to Rs 10,000 to a team at the juror's own discretion; there was no fixed prize pool, and the allocations were disbursements to the teams rather than the hypothetical per-submission willingness to fund scored above. A team's outcome is the total funding it raised, summed across its jurors. Jurors were not informed of schools' treatment status, or that the competition was part of a study.

B. An Incentivized Search Task

The Innovation Challenge measures what teams produced; this task measures how individual students search. March (1991) emphasizes the trade-off between exploiting known solutions and exploring new alternatives. This focus on the search process contrasts with the output-based definition of innovation underlying the Oslo Manual and the Innovation-S scale. Exploitation yields a modest but predictable payoff; exploration is high risk and high reward: an untried alternative may yield a better solution or nothing. The curriculum practices this search directly: in the ideate and make-and-test stages of each block, teams generate multiple candidate solutions, shortlist them against criteria, and build, test, and revise prototypes across sessions in which failure is expected (Section II.B).

To observe this search at the individual level, students in both arms completed the same incentivized task at endline, adapted from the experimentation game of Ederer and Manso (2013).

The student runs a street-food stand on a tablet for 12 rounds. Small stands of this kind are common across India, so the setting is familiar to students in both arms. The curriculum teaches exploration but does not teach business practices. A simulated business therefore allows the task to measure whether students

explore alternatives in a setting the curriculum did not cover.

Each round the student chooses a stand type (fresh juice, pani puri, or ice cream), a location (market, homes, or school), working hours (6 to 15), the number of menu items (3 to 12), and a selling price (Rs 11 to 20), a space of 9,000 configurations. Each stand has a maximum hourly profit, earned at its optimum, which the student does not know; profit in a round falls short of that maximum by a fixed penalty per unit of distance between the chosen inputs and the stand's optimal inputs, and is floored at zero. Formally, profit under stand s in round t is

$$\pi_s(\mathbf{c}_t) = \max\left\{0,\ V_s - \lambda_s \sum_d |c_{td} - c^*_{sd}|\right\}, \tag{1}$$

where V_s is the stand's maximum profit, λ_s its penalty rate, and $\mathbf{c}^*_s$ its optimal inputs. Distance is summed over the four non-stand dimensions; working hours, menu items, and selling price each lie on ten-point grids and location on a three-point grid, so a one-step deviation is penalized equally on every dimension. Profit is positive if and only if the summed distance is less than V_s/λ_s.

The stands differ in both parameters. Fresh juice has the lowest maximum and the smallest penalty ($V = 120$, $\lambda = 10$), pani puri is intermediate in both ($V = 180$, $\lambda = 20$), and ice cream has the highest maximum and the steepest penalty ($V = 250$, $\lambda = 30$). Higher maximum profits therefore coincide with a higher likelihood of earning nothing: a randomly chosen configuration returns zero profit on 25.7 percent of fresh-juice configurations, 54.0 percent of pani-puri configurations, and 63.3 percent of ice-cream configurations (Appendix Table D.1).

None of these parameters is disclosed to students. In the first round the screen shows a worked example in which a stand owner earned Rs 30 at the fresh-juice stand, the least remunerative and least risky of the three, so every student begins from a common reference point, a demonstrated configuration with an observed profit. Beyond the worked example, a student learns about the payoff structure only through the profit their own choices return each round. From the second round onward the screen shows the student's own best configuration and profit so far, so students can compare each new attempt with their highest profit so far. After each round, a summary table displays the choices and profits from all completed rounds. Students select all inputs in the first round; subsequent rounds retain the previous selections, allowing students to repeat the configuration, vary a single input, or vary multiple inputs simultaneously. Disengagement therefore registers as the absence of search rather than as noise. The announced prize, a geometry box, is awarded for the highest hourly profit discovered in the class.

Three design choices make the task a measure of exploratory behavior. First, before receiving feedback from any choices of their own, students decide whether to remain with the demonstrated stand or choose one

of the two unfamiliar alternatives. Because this round-one decision is made from identical task information and before students' experiences within the task can diverge, it records whether the student initiates search beyond a known workable solution. Initiation is therefore measured by whether the first-round stand choice departs from the demonstrated stand.

Second, the objective rewards the best solution the student discovers.[13] Once a student has found a profitable configuration, that profit remains the student's recorded maximum: an unsuccessful attempt cannot reduce the best profit already achieved. The objective therefore protects previously discovered value while the student experiments, and thereby tolerates the early failure emphasized by Manso (2011). It also gives the student reason to try unfamiliar alternatives rather than accumulate earnings by repeating a known configuration. Breadth captures the number of distinct alternatives explored, so I define it using the categorical choices of stand type and location rather than the quantitative variables of hours, menu items, and price. It is measured by the number of distinct stand-and-location combinations tried across the 12 rounds.

Third, the task measures whether students discover solutions that improve on their previous best and how closely they approach the chosen stand's optimum. Payoffs are deterministic: the same configuration always returns the same profit. Repeating a configuration therefore does not extend the search and produces the same result, whereas changing it reveals whether the search has produced a better solution. Because rounds 2–12 carry the preceding inputs forward, a student can vary one dimension while holding the others fixed and observe whether that local change raises or lowers profit. A zero-profit round reveals that the configuration lies outside the stand's profitable region but not which input should change, so finding a better solution requires continued experimentation. Moreover, each new attempt is evaluated against the student's displayed personal best. The round-level record therefore captures not only how broadly a student searches but whether the search becomes more successful: Discovery counts the rounds in which profit exceeds the student's own previous best, and Convergence is the closest approach to the chosen stand's optimum across the twelve rounds, an optimum the student never observes.

Initiation, Breadth, Discovery, and Convergence measure four aspects of search: whether the student departs from a demonstrated solution, how widely they range, how often they improve on their own best, and how close they come to the optimum. Because the payoff surface is known exactly, that closest approach is computed exactly rather than estimated. An exploration index summarizes the four margins: their equal-weighted mean, with each component standardized to the control group and distance entering with reversed sign.

[13]The baseline version of the task had used Ederer and Manso's original objective, cumulative earnings, which proved hard for students of this age to follow; a single best round is a goal a grade 9 student in this context could pursue.

The baseline version of the task followed the lemonade-stand design of Ederer and Manso (2013): the student ran one lemonade stand and chose its location, ingredients, and price. It enters the endline specifications as a pre-treatment control. Appendix D reports the remaining round-level process measures and the full parameterization.

C. The Remaining Skill Tasks

The remaining assessments measure emotional intelligence, fluid intelligence, creativity, and structured problem solving. Table 3 pairs each skill with its instrument. These assessments draw on established batteries and published experimental designs, adapted to the setting and piloted with out-of-sample students. Appendix D reports their adaptation, scoring, and reliability.

The find and learn stages practice perspective-taking and pattern recognition. I assess these skills through tests of emotional and fluid intelligence. Emotional intelligence is measured with the Reading the Mind in the Eyes Test (RMET) (Baron-Cohen et al. 2001), which shows the student a photograph of the eye region of a face and asks which of four emotions best describes what the person is feeling or thinking; the original stimuli were replaced with photographs appropriate to the Indian context, and the endline form has 11 items. Weidmann and Deming (2021) and Alan and Kubilay (2025) have also used the RMET to measure social and emotional intelligence. Fluid intelligence is measured with 12 matrix-reasoning items from the International Cognitive Ability Resource, a public-domain battery validated against standard intelligence tests (Condon and Revelle 2014): each item presents a three-by-three grid of geometric patterns with one cell missing, which the student completes from four options. Appendix Figures D.1 and D.2 show a sample item from each test.

The ideate stage practices divergent thinking and evaluative judgment. Three written tasks adapted from the PISA creative-thinking assessment measure idea generation, idea evaluation, and written expression (OECD 2024). In the idea-generation task, students propose three improvements to a standard bicycle, each scored for novelty. Idea evaluation asks the student to divide Rs 1,000 between two inventions, a torch with two bulbs and a wheelchair that folds into crutches, and to justify the allocation. The justification is scored according to whether it addresses cost, usefulness, or user needs rather than merely stating a preference. In the written-expression task, students write a short story about an image. Responses are scored for narrative structure and use of the image's elements. Responses are open-ended text in English or Telugu, scored against task-specific rubrics by a large language model. A research assistant blind to treatment status verified all assigned scores and confirmed more than 98 percent without changes. Appendix Table D.4 reports each task's rubric with example responses at every score level.

TABLE 3—CURRICULUM SKILLS, PRACTICE, AND STANDARDIZED MEASURES

Skill	Curriculum location	How the curriculum practiced it	Standardized measure (source)
Creative confidence	Ideate; Make and test; independent design	Generated multiple solutions, attempted solutions before seeing examples, built mock-ups, revised after feedback, and progressed from guided tasks to open-ended problem discovery.	Not separately assessed
Critical thinking	Find; Learn; Ideate; critique and selection	Distinguished problems from non-problems, investigated causes and stakeholders, examined design changes, compared alternatives, and evaluated ideas against explicit criteria.	ICAR matrix reasoning (Condon and Revelle 2014); idea evaluation, PISA (OECD 2024)
Creativity	Ideate; redesign challenges; sensor applications	Brainstormed under time and material constraints, redesigned objects, combined objects and features, and developed uses for sensors and outputs.	Idea generation and written expression, PISA (OECD 2024)
Empathy	Find; user-focused design; testing	Observed and interviewed users, identified their problems, designed for specified users in semi-open challenges, and gathered feedback on prototypes.	Reading the Mind in the Eyes Test (Baron-Cohen et al. 2001)
Problem solving	Full five-stage cycle	Defined and investigated problems, generated and selected solutions, planned resources, and built, tested, and revised prototypes with increasing independence.	4×4 Sudoku, difficulty-weighted (Alan, Boneva, and Ertac 2019); incentivized search task (Ederer and Manso 2013)
Collaboration	Team work; co-team feedback; peer review	Worked in fixed teams, divided tasks, compared ideas, exchanged feedback, and participated in class and showcase discussions, facilitated by student leaders.	Not separately assessed
Presentation	Present; Pause; Exhibition/showcase	Presented problems, evidence, solutions, prototypes, and feedback to class and school audiences using presentation cards and reflection prompts.	Recorded pitch (Section IV.A)
Communication	Interviews; peer feedback; presentations	Questioned users, articulated problems and design choices, listened to alternative perspectives, and explained prototypes and revisions.	Recorded pitch (Section IV.A)

Notes: The first three columns describe the curriculum's intended skills and practices, as mapped by its developer, Inqui-Lab. The final column lists the assessments selected by the research team, drawn from psychology (RMET and ICAR matrix reasoning), education (PISA creative-thinking tasks), and economics (Sudoku and the incentivized search task). The curriculum mapping draws on Inqui-Lab's product note, framework, session plans, student workbooks, portfolios, student-leader handbook, and Year-1 observation materials.

The make-and-test stage practices structured problem solving. I assess this skill with a five-round Sudoku task using four-by-four puzzles. A Sudoku puzzle is a well-defined problem in the sense of Newell and Simon (1972): the rules and goal are specified, and each step can be checked against the rules. The task requires little subject-specific knowledge, and the reduced grid makes it accessible to students across the range of academic attainment.

The measure adapts the incentivized easy-versus-hard choice design of Alan, Boneva, and Ertac (2019), with a Sudoku puzzle, solved by logic rather than arithmetic, in place of their number-matching real-effort task. Before each round the student chooses between an easy and a hard puzzle. One round is selected at the end to determine the prize. Students receive a gel pen for solving a hard puzzle, a ballpoint pen for solving an easy puzzle, and no prize for an unsolved puzzle. Choosing the hard puzzle entails a trade-off between a larger prize and a higher probability of failure. In the first round, half of the students, selected at random, played the hard puzzle regardless of their choice; from the second round onward every student played the puzzle they chose. Problem solving is the difficulty-weighted average score across the five rounds.

Challenge-seeking is the share of rounds 2–5 in which the student chose the harder puzzle. Because students receive no performance feedback between rounds, this measure captures willingness to attempt a harder problem rather than persistence following failure.

The matrix-reasoning test and the RMET were also administered at baseline; the creativity and Sudoku tasks were administered only at endline.

D. Academic Outcomes

Academic achievement is measured using endline tests and the grade 10 state graduation examinations. At endline, students completed 22 mathematics items and 20 science items on the tablet, alongside the skill tasks. I developed both tests using state textbooks, drawing on the grade 8 and grade 9 textbooks for mathematics and science. The mathematics test also incorporated a few items adapted from the endline assessments used by Muralidharan, Singh, and Ganimian (2019). Both tests were piloted in two out-of-sample schools.

The endline assessment also collected self-reported time use and school attendance. Students reported hours spent on five activities during a typical school week, which I express as shares of their total reported time. School attendance is calculated from the number of days students reported being absent during the preceding twelve school days.

The grade 10 board examinations took place one year after the program ended. Each of the six subjects (mathematics, science, social studies, and three language papers) is graded out of 100 marks: 80 from an externally proctored examination and 20 from a school-graded internal assessment. I use only the external component to avoid variation in school-based grading. Subject scores and the total are standardized to the control group.

V. Empirical Strategy

Treatment was assigned at the school level. I estimate treatment effects by comparing student and team outcomes across treatment and control schools. For outcome Y_{is} of student i in school s, I estimate

$$Y_{is} = \alpha + \beta\, T_s + R_s'\mu + X_i'\gamma + W_s'\theta + \rho B_i + \delta M_i + \varepsilon_{is}, \tag{2}$$

where T_s indicates a treatment school and R_s contains indicators for the three characteristics defining the randomization strata: tribal-welfare management, girls-only composition, and rural location (Section III.A). The vector X_i contains the student's age, household SES index, and household computer literacy, with indicators for imputed values; W_s contains the pre-treatment school characteristics listed in Appendix

Table A.2. The variable B_i is the student's baseline measure for the corresponding outcome family, and M_i indicates students who missed the baseline assessment. The coefficient β is the intent-to-treat effect of assignment to the curriculum. Standard errors are clustered at the school level, the unit of assignment.

Baseline controls improve precision. For each outcome, I use the closest pre-treatment measure of its construct, such as baseline mathematics for endline mathematics and the baseline exploration index for the search task. Table notes identify the baseline control for each outcome. Missing baseline scores are imputed at the school-section mean, allowing these students to remain in the estimation sample.

For team-level idea ratings, I estimate

$$Y_{gs} = \alpha + \beta\, T_s + R_s'\mu + W_s'\theta + \bar{X}_g'\phi + \varepsilon_{gs}, \tag{3}$$

where Y_{gs} is the rating of team g in school s. The vector $\bar{X}_g$ replaces the individual controls in equation (2) with team averages of members' baseline academic achievement, exploration index, fluid intelligence, emotional intelligence, household SES index, computer literacy, and age. It also includes team size and the numbers of members with baseline records and without a matched student record. Standard errors are clustered at the school level.

VI. Results

Participation in the Innovation Challenge. Treatment increased participation in an innovation competition offered on the same terms to students in both arms (Table 4, row 1). Among students assessed on submission day, participation rose by 8.0 percentage points from a control mean of 73.3 percent, significant at the 1 percent level. This corresponds to a roughly 30 percent reduction in nonparticipation.

Because submission-day attendance is a post-treatment outcome, I also estimate participation using the baseline-assessed roster. Treatment increased participation in this sample by 5.5 percentage points from a control mean of 56.4 percent ($p = 0.031$). This denominator is fixed before treatment: students named on submitted team rosters count as participants regardless of submission-day attendance, while students who left the school system remain in the denominator as nonparticipants. Both estimates use submitted team rosters to identify participation (Section IV.A).

Endline presence itself did not detectably differ by arm (a difference of 2.4 percentage points, $p = 0.29$). Appendix Table A.3 reports coverage at every stage and the differential-attrition tests.

TABLE 4—EFFECTS ON INNOVATION CHALLENGE PARTICIPATION AND SUBMISSION-ADJUSTED IDEA RATINGS

Outcome	Treatment effect	SE	Sharpened q	N
Participated in the Innovation Challenge (pp)	8.0***	(2.1)	[0.002]	4,388
Innovation-S novelty subscale (SD)	0.187***	(0.054)	[0.002]	1,493
Innovation-S value-add subscale (SD)	0.146***	(0.052)	[0.005]	1,493
Feasibility rating (SD)	0.109**	(0.047)	[0.009]	1,493
Evaluator willingness to fund (SD)	0.184***	(0.062)	[0.004]	1,493

Notes: Row 1 reports participation among 4,388 students assessed at endline. Participation equals one if a student is named on a submitted team roster; the effect is in percentage points (control mean: 73.3 percent). Rows 2–5 report submission-adjusted idea ratings for 1,493 team units: 1,216 scored submitting teams with at least one endline-assessed member and 277 constructed teams containing the 1,002 remaining students. Constructed teams contain up to four students from the same school and receive the no-submission score defined in Section VI. Rating effects are in standard deviations of the control group in this augmented sample. Appendix Table A.5 reports estimates conditional on submission.

Student and team controls follow equations (2) and (3); participation uses baseline mathematics and science as its baseline measures. Rating estimates average over 1,000 random groupings of non-submitters; their variances combine the average school-clustered variance and the between-grouping variance. Participation standard errors are clustered by school. Stars test each effect against zero using the clustered p-value. Bracketed values are Anderson sharpened false-discovery-rate q-values for the five outcomes. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Idea Quality. Treatment increased submission-adjusted ratings of novelty by 0.187 control-group standard deviations, value-add by 0.146, feasibility by 0.109, and willingness to fund by 0.184 (Table 4, rows 2–5). All four effects are significant at the 5 percent level, and all but feasibility at the 1 percent level.

The estimates use team units constructed from the endline-assessed sample. They include non-submitters because treatment can affect both submission and idea quality. Non-submission receives a raw score of zero on each rating, including Rs 0 for willingness to fund, before standardization; zero therefore represents no submitted idea rather than the control-group mean. Students on submitted teams enter through their team's evaluator ratings. Students not named on a submitted roster are grouped into units of up to four, matching the competition's prescribed team size (Section III.B). To account for the arbitrary grouping of non-submitters, I repeat the grouping 1,000 times and estimate the team-level specification (Section V) each time. I report the average effect, with a standard error that incorporates variation across groupings.

Appendix Figure A.1 shows the distribution of idea ratings: at every quality threshold from the no-submission margin through the middle of the distribution, the share of teams clearing the threshold is higher in treatment schools, with the gap narrowing only in the far upper tail. The distributional gains therefore extend beyond the participation margin.

Among submitted teams, treatment-school ideas also received higher ratings (Appendix Table A.5). The estimates are 0.265 standard deviations for novelty, 0.202 for value-add, and 0.206 for willingness to fund; the feasibility estimate is 0.100 and is not statistically significant. Because submission is affected by treatment,

this comparison is conditional on participation. Submitting teams do not differ significantly in their average baseline characteristics (joint $p = 0.79$). Treatment teams' baseline academic and skill scores are 0.02 to 0.07 standard deviations lower, with no statistically significant differences. Appendix Table A.4 reports these comparisons alongside team size and within-team skill dispersion. Lee bounds accounting for selection into submission remain positive for all ratings except feasibility; the bound endpoints are point estimates rather than confidence limits.

Jury Funding at the Shark Tank Finals. At the finals, professional jurors allocated actual funding among selected teams, providing a second assessment of their ideas. These jurors were distinct from the blind evaluators whose hypothetical willingness-to-fund ratings appear in Table 4. The 40 highest-rated teams from each arm were invited to pitch live, with funding disbursed at the jurors' discretion (Section IV.A). Figure 3 shows total funding raised by each of the 80 finalist teams, summed across jurors; Appendix Figure C.3 shows finalist teams presenting to jury panels at the event.

On average, treatment teams raised Rs 1,709 more than control teams, a difference of 27 percent, but neither the mean difference (randomization inference $p = 0.28$) nor a rank-based comparison of the full distributions (rank-sum $p = 0.42$) is statistically significant. The clearest divergence is in the upper tail: of the 10 teams that raised the most money, 8 were treatment teams (one-sided randomization inference $p = 0.075$), though the single highest-funded team was a control team. The jurors also scored each pitch on the same Innovation-S rubric before allocating funding, and their rubric scores were strongly correlated with their funding allocations (within-juror $r = 0.77$). The finals compare the 40 highest-rated teams from each arm. They therefore assess whether the best treatment teams outperform the best control teams in attracting jury funding, rather than whether treatment raised idea quality overall. With 80 teams, this comparison has low statistical power, so I report it as directional evidence. Appendix Table A.6 reports these estimates together with a robustness check comparing teams only within jury panels.

Innovation-Relevant Skills. Treatment improved students' exploration on a task unrelated to the products they developed for the Challenge (Table 5, Panel A). Panel A reports treatment effects on four measures of search behavior and on the exploration index, their equal-weighted average. Treatment raised the index by 0.094 index points, significant at the 1 percent level.

Treated students were more likely to leave the demonstrated stand and more successful at improving on previous attempts. Treated students were 5.6 percentage points more likely to leave the demonstrated stand in the first round (control mean 57.9 percent), a decision made before feedback from their own choices and

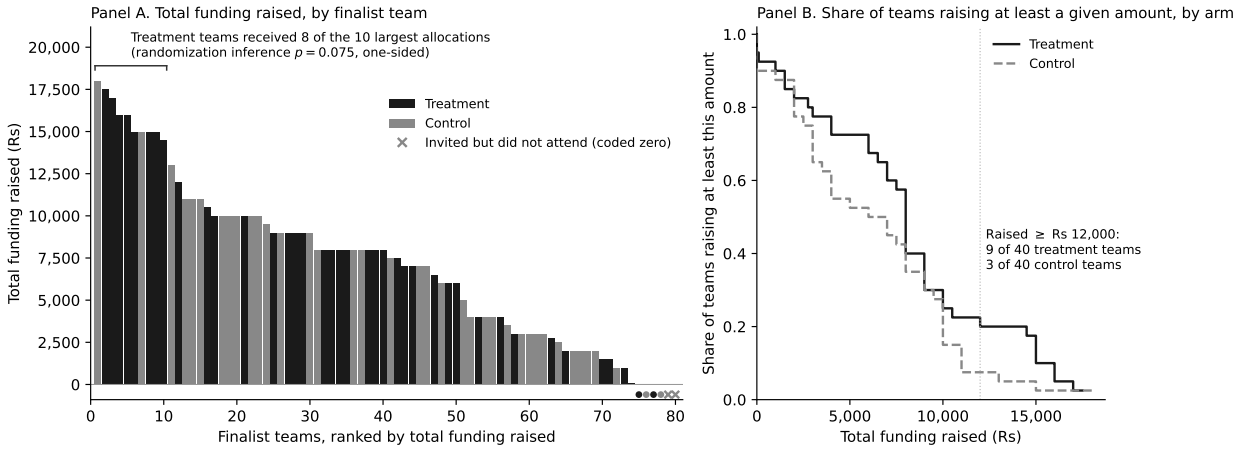


FIGURE 3. JURY FUNDING AT THE SHARK TANK FINALS

Notes: The sample is the 80 finalist teams of the Innovation Challenge, the top 40 teams from each arm ranked on the evaluators' ratings, invited to the Shark Tank-style finals. The outcome is the total funding a team raised from its jurors, in rupees. Panel A ranks all 80 teams by total funding raised; the two invited control teams that did not attend are coded as raising zero and marked with crosses. Panel B plots, for each arm, the share of teams that raised at least a given amount. Randomization inference permutes treatment assignment at the school level within randomization strata (5,000 draws) among the 78 attended teams, whose schools are observed. Estimates and robustness checks appear in Appendix Table A.6.

therefore on the same information in both arms. Over the remaining rounds, they improved on their own best profit in 0.101 additional rounds (control mean 1.26), and their minimum distance from the optimum was 0.256 lower (control mean 6.30); both effects are significant at the 5 percent level. The number of distinct stand-and-location combinations tried rose by 0.161 on a control mean of 4.50, significant at the 10 percent level. All five estimates survive the false-discovery-rate adjustment shown in brackets. Appendix Table D.2 reports search volume, responses to failure, and discovery and convergence estimates holding search volume fixed.

The largest standardized skill gain was in idea generation on an individual task. This score, based on the novelty of three proposed improvements to a bicycle, rose by 0.314 standard deviations, significant at the 1 percent level (Panel B). The estimated effects on the other two tasks were smaller. Idea evaluation rose 0.100 standard deviations, significant at the 10 percent level but not after adjustment for multiple hypotheses; the effect on written expression is 0.040 and statistically indistinguishable from zero.

Gains also extended to fluid intelligence and problem solving (Panel C). Fluid intelligence rose 0.111 standard deviations, and the difficulty-weighted Sudoku score rose 0.142 standard deviations. Both effects are significant at the 5 percent level and survive adjustment for multiple hypotheses.

Treatment had no detectable effects on emotional intelligence or challenge-seeking (Panel D). The effect on emotional intelligence, measured by the Reading the Mind in the Eyes Test, is 0.031 standard deviations; the 95 percent confidence interval rules out gains larger than 0.13 standard deviations. The effect

TABLE 5—EFFECTS ON INNOVATION-RELEVANT SKILLS

Outcome	Treatment effect	SE	Sharpened q	N
Panel A. Search behavior in the incentivized search task				
Exploration index (points)	0.094***	(0.025)	[0.002]	4,388
Initiation: left the demonstrated stand, round 1 (pp)	5.6***	(1.8)	[0.006]	4,388
Breadth: unique stand-location combinations tried	0.161*	(0.092)	[0.035]	4,388
Discovery: rounds improving on the personal best	0.101**	(0.039)	[0.012]	4,388
Convergence: minimum distance from the optimum	-0.256**	(0.115)	[0.020]	4,388
Panel B. Creativity				
Idea generation (SD)	0.314***	(0.078)	[0.001]	4,388
Idea evaluation (SD)	0.100*	(0.060)	[0.109]	4,388
Written expression (SD)	0.040	(0.065)	[0.219]	4,388
Panel C. Analytical reasoning				
Fluid intelligence (SD)	0.111**	(0.050)	[0.032]	4,388
Problem solving (SD)	0.142**	(0.064)	[0.032]	4,388
Panel D. Emotional intelligence and challenge-seeking				
Emotional intelligence (SD)	0.031	(0.048)	[1.000]	4,388
Challenge-seeking (SD)	0.019	(0.078)	[1.000]	4,388

Notes: The table reports treatment effects on innovation-relevant skills (Sections IV.B and IV.C), estimated at the student level on the 4,388 students who completed the endline assessment. Initiation is reported in percentage points; Breadth, Discovery, and Convergence are in natural units. The exploration index is the equal-weighted mean of the four control-standardized components, with Convergence reverse-signed. Outcomes in Panels B–D are standardized to have control-group mean zero and standard deviation one. **Panel A** reports search behavior in the incentivized search task; the outcomes are the four margins of a search and their exploration index. Initiation equals one if the student's round-1 stand differs from the demonstrated stand (control mean 57.9 percent); Breadth counts distinct stand-and-location combinations (control mean 4.50 of 9); Discovery counts rounds improving on the personal best (control mean 1.26); Convergence is the minimum distance from the chosen stand's optimum (control mean 6.30). **Panel B** reports three PISA-adapted written tasks scored by rubric: idea generation (novelty of three proposed improvements to a bicycle), idea evaluation (the justification of a budget allocation between two inventions), and written expression (a short story). **Panel C.** Fluid intelligence is a two-parameter item-response score from twelve ICAR items. Problem solving is the standardized difficulty-weighted mean score across five Sudoku rounds. **Panel D** reports emotional intelligence, from the eleven-item Reading the Mind in the Eyes Test (two-parameter item-response model), and challenge-seeking, the share of rounds 2–5 in which the student chose the harder puzzle in the Sudoku task.

Every regression controls for randomization strata, school infrastructure, and individual covariates (age, socioeconomic index, computer skills), together with the panel's pre-treatment baseline measure and a missing-value indicator: the baseline exploration index for Panels A and B, baseline fluid intelligence for Panel C, and, in Panel D, the baseline eyes test for emotional intelligence and baseline conscientiousness for challenge-seeking. Standard errors, in parentheses, are clustered at the school level, the level of treatment assignment; significance stars refer to the clustered p-value. Values in square brackets are Anderson sharpened false-discovery-rate q-values, computed within each panel family. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

on challenge-seeking, the willingness to attempt the harder of two puzzles in the Sudoku task, is 0.019 standard deviations, and treated and control students chose the difficult puzzle at the same rate in every round. Exploration improved without a detectable change in challenge-seeking on the Sudoku task. The problem-solving gain is similar under alternative scoring methods that remove the difficulty weighting or adjust for difficulty choice (Appendix Table A.7).

Academic Outcomes. The program produced no detectable declines in academic achievement at endline or on the grade 10 board examinations (Table 6). Sessions were held during after-school study or other non-academic periods, and the curriculum did not teach examination subjects. I therefore examine whether time devoted to the program came at a cost to academic achievement. Panel A reports the endline academic assessment. The effect on mathematics is -0.007 control-group standard deviations and the effect on science is 0.027; both are indistinguishable from zero, and the 95 percent confidence intervals rule out losses larger than 0.14 standard deviations in mathematics and 0.10 in science.[14]

Reported homework time shares and recent school attendance also showed no detectable declines. The reported share of weekly time spent on homework and study was 0.4 percentage points lower on a control mean of 19.2 percent ($p = 0.42$), and school attendance in the two weeks before the assessment was 1.4 percentage points higher in treatment schools on a control mean of 89.9 percent ($p = 0.15$). Reported weekly time shares were 1.3 percentage points higher for the Innovation Challenge ($p = 0.001$), 0.8 points higher for sports ($p = 0.04$), and 1.5 points lower for other activities ($p = 0.047$), which students described as including dance, singing, drawing, and reading. The share devoted to school periods showed no detectable decline (Appendix Table E.1).

Panel B reports effects on the grade 10 board examinations taken in March 2025, one year after the program ended. Administrative records cover 99.8 percent of the study cohort, including students not assessed at endline (Appendix Table A.1). Scores are the externally graded exam component, standardized to the control group. Point estimates are positive in five of the six subjects and never negative by more than 0.006 standard deviations. The effect is 0.112 on mathematics ($p = 0.45$), -0.006 on science ($p = 0.96$), and 0.135 on the total score ($p = 0.24$); the pass rate is 0.6 percentage points higher on a control mean of 96.7 percent ($p = 0.44$). Estimated effects on the three language papers range from 0.149 to 0.199 standard deviations, each significant at the 10 percent level. None survives the false-discovery-rate adjustment ($q \geq 0.34$), so I treat them as suggestive evidence of language gains rather than a confirmed effect. Across the six subjects, the confidence intervals rule out losses larger than 0.24 standard deviations in any single paper and 0.09 on the total score.

Ceiling compression limits the interpretation of these bounds. Among control students, 18 percent of mathematics and 13 percent of science external scores fall within 10 percent of the 80-mark maximum, and the correlation between the board score and the endline assessment is 0.20 in mathematics and 0.30

[14] A midline assessment in February 2023, near the end of Year 1 and in the first full academic year after COVID-related school closures, showed a pattern that did not persist. Mathematics was 0.137 standard deviations lower ($p = 0.007$), science was unchanged (0.000), and emotional intelligence was marginally lower (-0.101, $p = 0.051$; $N = 4{,}620$, estimated with the Table 6 specification). The mathematics gap was not detected at endline (Panel A) or on the board examinations (Panel B). Appendix F compiles students' open-ended feedback from the same assessment.

TABLE 6—EFFECTS ON ACADEMIC OUTCOMES

Outcome	Treatment effect	SE	Sharpened q	N
Panel A. Endline assessment				
Mathematics	-0.007	(0.067)	[1.000]	4,388
Science	0.027	(0.061)	[1.000]	4,388
Time on homework (% of reported time)	-0.36	(0.45)	[1.000]	4,388
School attendance, past two weeks (%)	1.44	(1.00)	[1.000]	4,388
Panel B. Grade 10 board examinations				
Mathematics	0.112	(0.147)	[0.598]	5,632
Science	-0.006	(0.119)	[0.945]	5,632
Social studies	0.023	(0.093)	[0.945]	5,632
First language	0.149*	(0.088)	[0.338]	5,632
Second language	0.199*	(0.106)	[0.338]	5,631
Third language	0.153*	(0.086)	[0.338]	5,631
Total score	0.135	(0.115)	[0.436]	5,633
Passed (pp)	0.64	(0.83)	[0.598]	5,642

Notes: The table reports treatment effects on academic outcomes at two horizons. **Panel A** reports the endline academic assessment, administered in the same battery as the skill tasks in Table 5, for the 4,388 students who completed the endline. Mathematics and science are item-response scores standardized to the control group. Time on homework is the share of the student's reported weekly time, in percentage points (control mean 19.2 percent), from a five-activity weekly hours grid; grids totaling 80 or more hours per week (28.3 percent of students) had all five activity shares replaced by the rounded mean shares of unflagged classmates in the same school section, and this replacement flag is balanced across arms ($p = 0.73$). School attendance is the percent of the twelve school days in the two weeks before the assessment on which the student was present, recoded from the item asking how many days the student was absent from school (control mean 89.9 percent). **Panel B** reports the March 2025 grade 10 board examinations, whose administrative records cover 99.8 percent of the 5,642-student study cohort. Subject scores are the externally graded examination component, standardized to the control group; the total score is the standardized total across subjects; Passed is an indicator scaled to percentage points (control mean 96.7 percent), with the nine students absent from or withheld at the examination coded as not passing. Every regression controls for randomization strata, school infrastructure, and individual covariates (age, socioeconomic index, computer skills), together with the row's pre-treatment baseline measure and its missing-value indicator: baseline mathematics for the mathematics rows, baseline science for the science rows, and both baseline scores for the remaining rows. Standard errors, in parentheses, are clustered at the school level, the level of treatment assignment; significance stars refer to the clustered p-value. Values in square brackets are Anderson sharpened false-discovery-rate q-values, computed within each panel. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

in science. The board examinations have limited power to detect effects among the strongest students; the endline assessments in Panel A provide complementary evidence less affected by this ceiling (Appendix Figure E.1). Singh (2024) documents similar compression in administrative learning data from younger grades.

Academic Scores and Innovation-Relevant Skills. Academic scores, including the grade 10 board scores that screen entry into science streams, weakly predict which students are strongest in the innovation-relevant skills. Figure 1 plots the four endline skill measures against the baseline academic score and the grade 10 board-examination total among control students. The baseline academic score explains at most 12 percent of the variation in idea generation, problem solving, fluid intelligence, or exploration, and the grade 10 board

score at most 10 percent. Both maxima are for problem solving ($N = 1{,}602$ control students with observed baseline scores and $N = 2{,}123$ with board scores). A screen limited to the top academic quintile would miss roughly 60 to 80 percent of the students in the top quintile of each skill.

Appendix Figure A.2 reports pairwise correlations among the academic and skill measures. Academic measures correlate 0.35 to 0.50 with one another; their correlations with the four skill measures range from 0.06 to 0.44. These comparisons are descriptive; they are not treatment effects.

Heterogeneity. Table 7 reports treatment-effect heterogeneity by pre-treatment characteristics. For each outcome in Tables 4–6, I estimate differences in treatment effects by school gender composition and baseline academic score. Students are split at the control-group median of the combined baseline mathematics and science score; teams are split at the median of team-average scores among control teams. Appendix Table A.8 reports the same outcomes by the two remaining stratification characteristics, tribal-welfare management and rural location, and Appendix Table A.9 reports the joint-interaction and within-subsample estimates.

Participation gains were largest in subgroups with lower control-group participation, while exploration gains were observed across subgroups. Treatment increased participation by 12.9 percentage points in boys' schools, where control-group participation was 64.9 percent, and by 3.7 points in girls' schools, where control-group participation was 80.5 percent ($p_{\text{diff}} = 0.031$). Participation increased by 10.2 percentage points among students below the baseline academic median and by 5.5 points among those above it. Both effects are statistically significant; the difference between the two effects is not statistically significant ($p_{\text{diff}} = 0.11$). Treatment increased rated idea quality among teams both below and above the baseline academic median. Estimated effects on novelty are 0.205 and 0.168 standard deviations, respectively, and the smallest interaction p-value across the four ratings is 0.30. The gains in Table 4 therefore extend to teams with lower baseline academic scores. Estimated effects on the exploration index range from 0.052 to 0.130 points across the subgroups in Table 7 and Appendix Table A.8, with no statistically significant differences between subgroups.

Idea-generation and fluid-intelligence gains were larger among students with higher baseline academic scores. Estimated effects on idea generation are 0.503 standard deviations in girls' schools and 0.096 in boys' schools ($p_{\text{diff}} = 0.011$). The corresponding estimates for idea evaluation are 0.218 and -0.034 standard deviations ($p_{\text{diff}} = 0.042$). Treatment effects were larger among students above the baseline academic median for idea generation (0.456 versus 0.200, $p_{\text{diff}} = 0.014$) and fluid intelligence (0.222 versus 0.020, $p_{\text{diff}} = 0.014$).

Students in girls' schools scored 0.36 standard deviations higher on the baseline academic assessment, so

TABLE 7—HETEROGENEOUS TREATMENT EFFECTS

	By school gender composition			By baseline academic score			
	Boys' schools	Girls' schools	Difference p	Below median	Above median	Difference p	N
Panel A. Innovation Challenge production (Table 4 outcomes)							
Participation (pp)	12.9***	3.7	0.031	10.2***	5.5**	0.107	4,388
	(3.1)	(2.8)		(2.5)	(2.6)		
Novelty (SD)	0.241***	0.140**	0.364	0.205***	0.168**	0.716	1,493
	(0.090)	(0.064)		(0.070)	(0.079)		
Value-add (SD)	0.222**	0.079	0.188	0.174**	0.116	0.563	1,493
	(0.086)	(0.063)		(0.068)	(0.077)		
Feasibility (SD)	0.180**	0.047	0.157	0.160**	0.054	0.303	1,493
	(0.076)	(0.056)		(0.069)	(0.069)		
Willingness to fund (SD)	0.236**	0.138*	0.440	0.204***	0.163*	0.719	1,493
	(0.097)	(0.080)		(0.076)	(0.091)		
Panel B. Innovation-relevant skills (Table 5 outcomes)							
Exploration index (points)	0.052	0.130***	0.130	0.091***	0.097***	0.890	4,388
	(0.035)	(0.036)		(0.031)	(0.035)		
Idea generation (SD)	0.096	0.503***	0.011	0.200***	0.456***	0.014	4,388
	(0.087)	(0.123)		(0.072)	(0.102)		
Idea evaluation (SD)	-0.034	0.218**	0.042	0.025	0.196***	0.037	4,388
	(0.073)	(0.093)		(0.061)	(0.073)		
Written expression (SD)	-0.007	0.081	0.504	0.003	0.090	0.264	4,388
	(0.094)	(0.090)		(0.071)	(0.069)		
Fluid intelligence (SD)	0.062	0.153**	0.339	0.020	0.222***	0.014	4,388
	(0.069)	(0.069)		(0.063)	(0.063)		
Problem solving (SD)	0.250**	0.046	0.164	0.158**	0.126	0.697	4,388
	(0.115)	(0.076)		(0.070)	(0.083)		
Emotional intelligence (SD)	0.021	0.039	0.848	-0.021	0.094	0.136	4,388
	(0.066)	(0.069)		(0.059)	(0.062)		
Challenge-seeking (SD)	0.054	-0.011	0.686	0.019	0.019	0.998	4,388
	(0.096)	(0.123)		(0.083)	(0.104)		
Panel C. Academic outcomes (Table 6 outcomes)							
Endline mathematics (SD)	-0.044	0.025	0.615	-0.112	0.118	0.009	4,388
	(0.097)	(0.094)		(0.069)	(0.087)		
Endline science (SD)	-0.034	0.079	0.358	-0.011	0.075	0.334	4,388
	(0.085)	(0.086)		(0.059)	(0.088)		
Grade 10 total score (SD)	0.014	0.242*	0.326	0.158	0.108	0.681	5,633
	(0.187)	(0.138)		(0.141)	(0.116)		

Notes: Participation and skill outcomes use 4,388 endline-assessed students; team ratings use 1,493 team units; the grade 10 row uses 5,633 students. Outcomes and units follow Tables 4–6. The first two columns of each block report subgroup effects. Stars test each effect against zero; the "Difference" columns report p-values for equality of the two effects.

Student and team-rating rows use interacted versions of equations (2) and (3), respectively, adding the group indicator and its interaction with treatment. The girls-only indicator is removed from the strata controls when it defines the groups. Individual academic groups use the control-group median of the combined 27-item baseline academic score among students with observed scores; missing scores are imputed as in Section V. Team-rating groups use the median of team-average baseline academic scores among control teams, recomputed within each of the 1,000 partitions of non-submitters. The remaining controls and baseline measures follow the corresponding results-table row. Standard errors, in parentheses, are clustered by school and use the full variance-covariance matrix for each subgroup effect; for team ratings, variances also incorporate between-partition variation. Appendix Tables A.8 and A.9 report the other stratification dimensions and robustness checks. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

I also estimate the gender and baseline-score interactions jointly.[15] For idea generation, the estimated gender interaction is 0.337 ($p = 0.020$), and the baseline-score interaction is 0.198 ($p = 0.049$). The fluid-intelligence gender interaction is not statistically significant (0.048, $p = 0.61$). The baseline-score interactions remain significant for both outcomes.[16]

Endline mathematics estimates are negative in the third baseline-score quintile (-0.148 standard deviations, $p = 0.099$) and positive in the top quintile (0.230, $p = 0.007$; Figure 4, Panel E). This pattern is not evident in grade 10 mathematics one year later (Panel F), although board scores are compressed near the top of the scale (Appendix Figure E.1).[17]

Exploration and idea-generation gains extended across the academic distribution (Figure 4). Estimated participation effects range from 9.7 to 12.0 percentage points in the bottom three quintiles. The estimated effect in the top quintile is 3.3 percentage points and is not statistically significant. Estimated effects on the exploration index range from 0.07 to 0.12 points across quintiles, with no clear gradient by baseline academic score. Treatment effects on idea generation are statistically significant in every quintile. The largest estimate is 0.64 standard deviations in the top quintile. Treatment increased problem-solving scores in the lowest and highest baseline academic quintiles.

Direction of Ideas. Appendix Figure A.3 describes the sectors and intended users of the 1,225 scored submissions by three pre-treatment team attributes: school gender composition, management society, and team baseline academic score. Sector composition differs by management society: submissions from tribal-welfare schools are more concentrated in agriculture and transport, while those from social-welfare schools are more concentrated in the environment and household life (joint $p = 0.003$, sharpened $q = 0.018$ across the figure's eight tests).

Sector composition also differs between girls' and boys' schools. Public welfare, health, and households account for 32.9 percent of submissions from girls' schools and 23.6 percent from boys' schools. Environment and transport account for 35.6 percent of submissions from boys' schools and 23.9 percent from girls' schools (Panel C; $p = 0.011$ and 0.004, $q = 0.022$ and 0.018). These aggregated sector groups reflect the pattern in Panel A. Joint tests comparing girls' and boys' schools across the full sector and intended-user classifications yield $p = 0.036$ and 0.076, respectively.

[15] All schools in the sample are single-gender, so the comparison is between boys' and girls' schools, not between boys and girls within a school; differences may reflect classroom composition, peers, or teacher characteristics that vary with school type. Appendix Table A.9 shows the idea-generation differential is concentrated in the girls' schools of the social-welfare system.

[16] Across the sixteen interaction tests within each dimension, the sharpened q-values are 0.079 for the baseline-score interactions in idea generation and fluid intelligence and 0.210 for the gender interaction in idea generation.

[17] Among students with observed baseline scores, estimated treatment effects in the top quintile are 0.62 standard deviations for idea generation (interaction $p = 0.011$) and 0.22 standard deviations for endline mathematics (interaction $p = 0.009$).

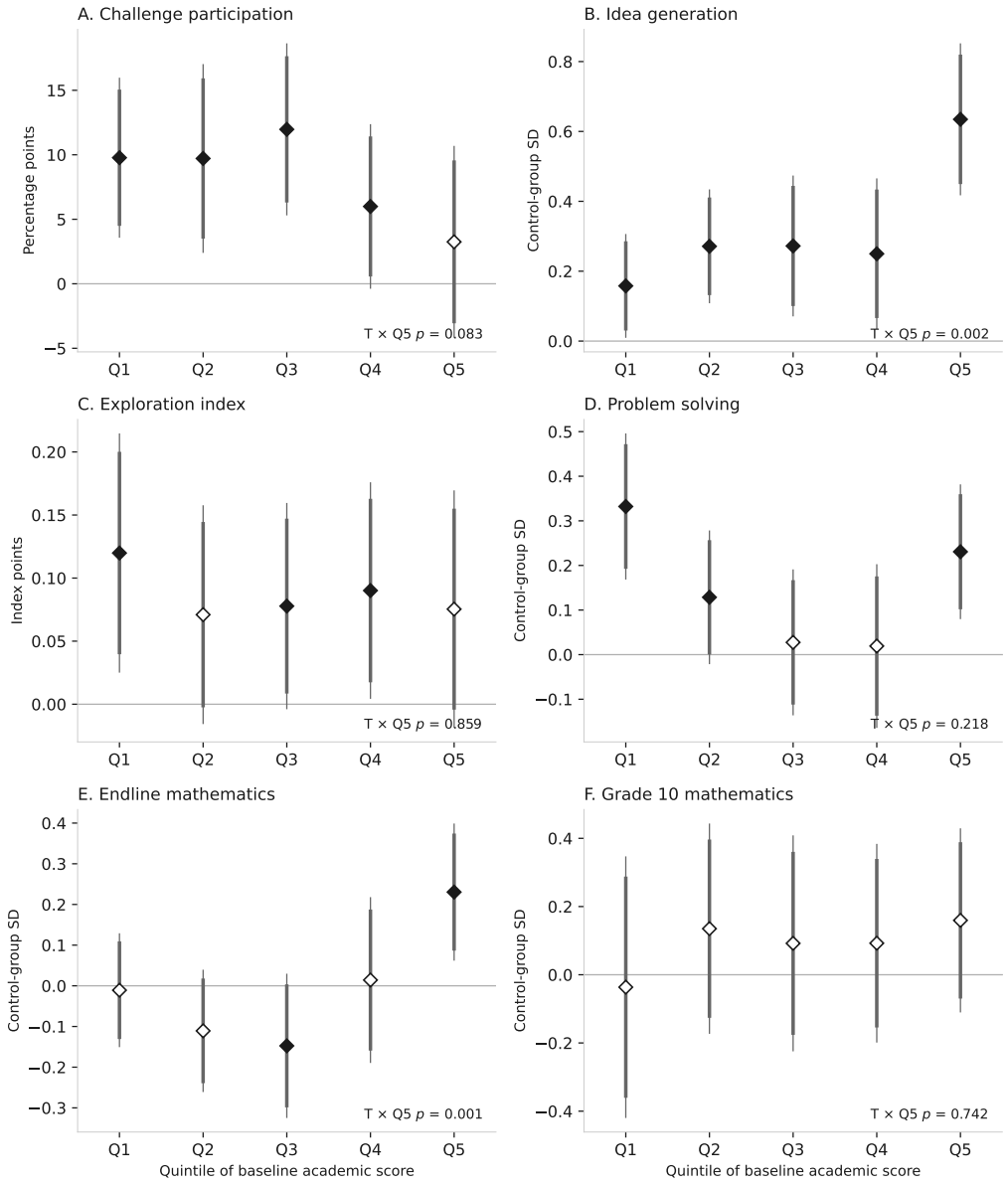


FIGURE 4. TREATMENT EFFECTS BY BASELINE ACADEMIC SCORE QUINTILE

Notes: Points show within-quintile treatment effects; thick bars are 90 percent and thin bars 95 percent confidence intervals from school-clustered standard errors. Filled markers indicate clustered $p < 0.10$; the solid gray line marks zero. Panels A–E use 4,388 endline-assessed students, including the search-task sample in Panel C; Panel F uses 5,632 students with grade 10 board mathematics records. Quintiles are formed within each estimation sample using the combined score from the 27 baseline mathematics and science items. Baseline academic scores are missing in the analysis data for 25.3 percent of students in Panels A–E and 28.6 percent in Panel F; these scores are imputed as in Section V. Panel A is in percentage points; Panel C is in exploration-index points, constructed as in Table 5, Panel A. Outcomes in Panels B and D–F have control-group mean zero and standard deviation one. Each regression includes randomization strata, school infrastructure, individual controls, and the outcome's baseline control with its missing-value indicator. Each panel reports the p-value on the full-sample interaction of treatment with a top-quintile indicator, comparing Q5 with Q1–Q4. Full-sample estimates appear in Tables 4–6.

Differences in the aggregated sector shares are not statistically significant by management society ($p \geq 0.79$) or between treatment and control submissions ($p \geq 0.49$). Comparisons by team baseline academic score yield $p = 0.17$ for public welfare, health, and households and $p = 0.09$ for environment and transport. The team academic score is also not significantly related to the full sector composition (joint $p = 0.52$) or to intended users ($p = 0.58$). Because teams form after treatment and submission is voluntary, these comparisons describe the composition of submitted ideas; they are not treatment effects.[18]

Robustness. Appendix Tables A.10 through A.12 re-estimate the competition, skill, academic test-score, and pass outcomes of Tables 4 through 6 under specifications ranging from treatment alone to a saturated stratum-cell fixed-effects model: no estimate changes sign, the idea-rating effects are larger without controls than with them, and replacing the three strata indicators with the eight stratum cells leaves every coefficient essentially unchanged. The fluid-intelligence and problem-solving estimates remain positive across specifications, although they are not statistically significant without covariate adjustment.

VII. Cost Effectiveness and Implications for Scale

The estimated financial cost of the two-year program was approximately $80 (Rs 6,600) per student, or about $40 per student per year, at the 40-school scale evaluated. The estimate covers classroom materials, student-leader training, delivery, monitoring, program management, and an allocation for curriculum development; the Innovation Challenge and its finals were organized by the government as part of its standing competition and added no delivery cost. Per treatment school, the program cost approximately $2,800 (Rs 2.3 lakh) per year.

Because effects were measured after two program years, cost-effectiveness is calculated using the full two-year cost of $80 per student. Across the four population-level measures of idea quality in Table 4, treatment effects of 0.109 to 0.187 standard deviations imply a cost of approximately $43 to $73 per 0.1 standard deviation of improvement in rated idea quality. A relevant fiscal benchmark is expenditure on India's school innovation infrastructure: the Atal Tinkering Labs program provides selected schools with approximately $24,000 (Rs 20 lakh) for laboratory establishment and operation over five years, about $4,800 per school per year (Atal Innovation Mission 2024), against $2,800 per year for this curriculum, which required no dedicated laboratory. The programs differ in inputs and student coverage, so this comparison indicates budgetary scale rather than relative cost-effectiveness.

[18]Within the submitted sample, joint tests comparing treatment and control submissions yield $p = 0.108$ for sector composition and $p = 0.003$ for intended-user composition (Appendix Figure A.3).

VIII. Discussion and Conclusion

This paper shows that schools can develop students' skills in generating and developing new ideas. In a randomized evaluation across 80 government residential schools in Telangana, India, a two-year curriculum increased participation in an innovation competition, improved the rated quality of student ideas, and developed innovation-relevant skills. Although the instructional sequence and materials were standardized, students chose the problems they addressed and the solutions they developed. The findings establish that structured instruction can improve open-ended production, providing one component of the exposure to innovation that families and neighborhoods supply unevenly (Bell et al. 2019).

The individual assessments provide evidence that learning extended beyond the teams' instructional projects. Students improved in idea generation, exploration, fluid intelligence, and well-defined problem solving. There were no detectable effects on written expression, emotional intelligence, or challenge-seeking.

The measurement approach makes this early production observable before occupational and market outcomes emerge. Independent evaluators, blind to treatment status, assessed student ideas using a rubric informed by the Oslo Manual and practitioner input. Professional jurors subsequently assessed finalists' pitches and allocated actual funding, providing a second assessment of the selected teams' ideas.

The common competition provides a policy-relevant comparison between access to an opportunity and preparation to use it. Both treatment and control schools received the same opportunity to submit ideas, but instruction increased participation and improved submissions. Governments operating school innovation competitions may therefore benefit from complementing these opportunities with guided practice. Participation, idea ratings, exploration, and idea generation improved among students and teams with below-median prior academic scores. Gains in idea generation and fluid intelligence were larger above the median. Broadening participation and complementarity with prior achievement can thus coexist.

The findings also bear on the use of academic performance to identify innovation-relevant talent. In India, grade 10 examination scores influence entry into science streams, and entrance examinations govern admission to technical degrees. The incentives surrounding these examinations can direct instruction toward examination performance rather than skills with broader applications (Muralidharan 2024). Among control students, academic scores explained at most 12 percent of variation in the assessed skills. Academic rankings therefore provide an incomplete account of the skills measured here. Developing these skills has a role both before academic selection and among students who advance through it.

The delivery model offers a concrete approach to advancing the creativity and innovation objectives of India's National Education Policy (Government of India 2020). Two student leaders per classroom received

20 hours of training and led sessions using structured workbooks, with teacher support and regular external monitoring. Specialist expertise entered through curriculum design, training, and continuing support, without requiring specialist classroom teachers or dedicated laboratories. At approximately $40 per student annually, the program produced these gains without a detectable average reduction in endline academic achievement or subsequent state graduation examination scores.

These findings suggest a role for schools in innovation policy, alongside investments in research infrastructure, finance, and entrepreneurship support. Preparing students to formulate and test ideas may improve their use of these later opportunities and eventually expand participation in the search for new ideas. Broader participation may also affect which problems receive attention. Girls' and boys' teams, and teams from tribal- and social-welfare schools, proposed ideas in different sectors. These descriptive differences are consistent with evidence linking inventors' backgrounds to the direction of invention (Bell et al. 2019; Koning, Samila, and Ferguson 2021; Einiö, Feng, and Jaravel 2025).

The innovation outcomes were measured at the end of instruction. Following students beyond school could test whether these gains persist and affect subsequent invention or entrepreneurship. Further work could examine how improvements in individual skills contribute to team production and whether the gains can be sustained with less external monitoring or alternative training arrangements. The experiment establishes an earlier link: schools can develop students' skills in generating and developing ideas and improve the quality of the ideas they produce.

REFERENCES

Agarwal, Ruchir, and Patrick Gaulé. 2020. "Invisible Geniuses: Could the Knowledge Frontier Advance Faster?" *American Economic Review: Insights* 2 (4): 409–424.

Agarwal, Ruchir, Patrick Gaulé, and Yuyan Jiang. 2025. "Finding Young Einsteins: Olympiads and STEM Talent Discovery." Global Talent Fund Working Paper WP/25/01.

Aghion, Philippe, Ufuk Akcigit, Ari Hyytinen, and Otto Toivanen. 2017. "The Social Origins of Inventors." NBER Working Paper 24110.

Alan, Sule, and Elif Kubilay. 2025. "Empowering Adolescents to Transform Schools: Lessons from a Behavioral Targeting." *American Economic Review* 115 (2): 365–407.

Alan, Sule, and Ipek Mumcu. 2024. "Nurturing Childhood Curiosity to Enhance Learning: Evidence from a Randomized Pedagogical Intervention." *American Economic Review* 114 (4): 1173–1210.

Alan, Sule, Teodora Boneva, and Seda Ertac. 2019. "Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit." *Quarterly Journal of Economics* 134 (3): 1121–1162.

Al-Ubaydli, Omar, John A. List, and Dana L. Suskind. 2017. "What Can We Learn from Experiments? Understanding the Threats to the Scalability of Experimental Results." *American Economic Review* 107 (5): 282–286.

Amaral, Sofia, Aakash Bhalothia, Ritam Chaurey, Isis Gaddis, Gaurav Khanna, Samreen Malik, Abhiroop Mukhopadhyay, Nishith Prakash, and Raghav Rakesh. 2025. "Building Entrepreneurial Mind-Sets: Evidence from a Large-Scale Educational Intervention in India." World Bank Group South Asia Gender Innovation Lab Policy Brief, Washington, DC. https://documents.worldbank.org/curated/en/099520405072521933.

Anderson, Lorin W., and David R. Krathwohl, eds. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* New York: Longman.

ASER Centre. 2025. "Annual Status of Education Report (Rural) 2024." ASER Centre, New Delhi. Provisional report, released January 28, 2025. https://asercentre.org/aser-2024/ (accessed September 26, 2026).

Asher, Sam, Paul Novosad, and Charlie Rafkin. 2024. "Intergenerational Mobility in India: New Measures and Estimates Across Time and Social Groups." *American Economic Journal: Applied Economics* 16 (2): 66–98.

Ashraf, Nava, Natalie Bau, Corinne Low, and Kathleen McGinn. 2020. "Negotiating a Better Future: How Interpersonal Skills Facilitate Intergenerational Investment." *Quarterly Journal of Economics* 135 (2): 1095–1151.

Ashraf, Nava, Natalie Bau, Corinne Low, and Xiaoyue Shan. 2025. "Ten Years of Relational Power: The Long-Run Effects of Teaching Negotiation Skills to Adolescent Girls." NBER Working Paper 34339. Conditionally accepted, *AER: Insights*.

Atal Innovation Mission. 2024. "ATL Guidebook: Integrating Innovation in School Education." Atal Innovation Mission, NITI Aayog. https://aim.gov.in/pdf/ATL-Guidebook.pdf (accessed September 26, 2026).

Bando, Rosangela, Emma Näslund-Hadley, and Paul Gertler. 2019. "Effect of Inquiry and Problem Based Pedagogy on Learning: Evidence from 10 Field Experiments in Four Countries." NBER Working Paper 26280.

Baron-Cohen, Simon, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb. 2001. "The "Reading the Mind in the Eyes" Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-Functioning Autism." *Journal of Child Psychology and Psychiatry* 42 (2): 241–251.

Bell, Alexander, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen. 2019. "Who Becomes an Inventor in America? The Importance of Exposure to Innovation." *Quarterly Journal of Economics* 134 (2): 647–713.

Bharti, Nitin Kumar. 2018. "Wealth Inequality, Class and Caste in India, 1961–2012." World Inequality Lab Working Paper 2018/14.

Blimpo, Moussa P., and Todd Pugatch. 2021. "Entrepreneurship Education and Teacher Training in Rwanda." *Journal of Development Economics* 149: 102583.

Bloom, Benjamin S., Max D. Engelhart, Edward J. Furst, Walker H. Hill, and David R. Krathwohl. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain.* New York: David McKay.

Bloom, Nicholas, John Van Reenen, and Heidi Williams. 2019. "A Toolkit of Policies to Promote Innovation." *Journal of Economic Perspectives* 33 (3): 163–184.

Brouillette, Jean-Félix. 2025. "The Lost Marie Curies and Foregone Economic Growth." *American Economic Journal: Macroeconomics*. Forthcoming.

Brown, Christina, Supreet Kaur, Geeta Kingdon, and Heather Schofield. 2025. "Cognitive Endurance as Human Capital." *Quarterly Journal of Economics* 140 (2): 943–1002.

CBSE. 2010. "Admission of CBSE Students in Class XI for Academic Year 2010–11 after Introduction of Grading at Secondary School Level by the Board." Central Board of Secondary Education, Circular No. 24/28-05-2010. https://www.cbse.gov.in/circulars/cir24-2010.pdf (accessed August 3, 2026).

CGD. 2025. "The Big Returns to Early Investments in Foundational Literacy and Numeracy." Center for Global Development event and transcript, October 17. https://www.cgdev.org/event/big-returns-early-investments-foundational-literacy-and-numeracy.

Chioda, Laura, Paul J. Gertler, David Contreras-Loya, and Dana Rose Carney. 2026. "Making Entrepreneurs: Long Term Returns to Training Youth in Business Skills." NBER Working Paper 34637.

Condon, David M., and William Revelle. 2014. "The International Cognitive Ability Resource: Development and Initial Validation of a Public-Domain Measure." *Intelligence* 43: 52–64.

Cunha, Flavio, and James Heckman. 2007. "The Technology of Skill Formation." *American Economic Review* 97 (2): 31–47.

Deccan Chronicle. 2026. "1.72 Lakh Students Appear for TG CET-2026." February 23. https://www.deccanchronicle.com/southern-states/telangana/172-lakh-students-appear-for-tg-cet-2026-1939173 (accessed September 26, 2026).

Dejuan, Irene, and Elisa Sanz Navarro. 2023. "Process Evaluation of the School Innovation Challenge Programme in Telangana, India." UNICEF Regional Office for South Asia. Prepared by the Center for Evaluation and Development (C4ED). https://evaluationreports.unicef.org/GetDocument?documentID=19421&fileID=51439 (accessed September 26, 2026).

Delhi DoE. 2026. "Guidelines for Parents of Children Seeking Admission in Government Schools through the Online Module in Class XI under Non-Plan Admissions for the Academic Session 2026–27." Government of National Capital Territory of Delhi, Directorate of Education, School Branch. https://edudel.nic.in/upload/upload_2025_26/SchoolBranch_GuidelinesForParentsOfChildrenSeekingAdmissionInGovtSchoolsThroughOnlineModuleInClass

XIunderNonPlanAdmissions2026_27_dt_23042026_180.PDF (accessed August 3, 2026).

Deming, David J. 2017. "The Growing Importance of Social Skills in the Labor Market." *Quarterly Journal of Economics* 132 (4): 1593–1640.

Deming, David J. 2022. "Four Facts about Human Capital." *Journal of Economic Perspectives* 36 (3): 75–102.

Ederer, Florian, and Gustavo Manso. 2013. "Is Pay for Performance Detrimental to Innovation?" *Management Science* 59 (7): 1496–1513.

Einiö, Elias, Josh Feng, and Xavier Jaravel. 2025. "Social Push and the Direction of Innovation." *American Economic Review*. Forthcoming.

Evans, David, and Susannah Hares. 2021. "Should Governments and Donors Prioritize Investments in Foundational Literacy and Numeracy?" Center for Global Development. May 4. https://www.cgdev.org/publication/should-governments-and-donors-prioritize-investments-foundational-literacy-and-numeracy.

Gneezy, Uri, and Jan Potters. 1997. "An Experiment on Risk Taking and Evaluation Periods." *Quarterly Journal of Economics* 112 (2): 631–645.

Government of India. 2020. "National Education Policy 2020." Ministry of Human Resource Development, New Delhi. https://static.pib.gov.in/WriteReadData/userfiles/NEP_Final_English_0.pdf.

Government of Telangana. 2026. "Telangana Gurukulam Common Entrance Test 2026: Results." https://tgcet.cgg.gov.in/ (accessed July 6, 2026).

Guilford, J. P. 1967. *The Nature of Human Intelligence.* New York: McGraw-Hill.

Gupta, Saloni. 2025. "Can Innovation Be Taught in Schools? Experimental Evidence from India." *AEA RCT Registry*, May 26. https://doi.org/10.1257/rct.9696.

Heckman, James J., and Tim Kautz. 2012. "Hard Evidence on Soft Skills." *Labour Economics* 19 (4): 451–464.

Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24 (3): 411–482.

Hoisl, Karin, Hans Christian Kongsted, and Myriam Mariani. 2023. "Lost Marie Curies: Parental Impact on the Probability of Becoming an Inventor." *Management Science* 69 (3): 1714–1738.

Inqui-Lab Foundation. 2026. "School Innovation Marathon 2025 Report." Forthcoming.

Jha, Anushka. 2025. "As Shark Tank India Turns Five, Sony Bets on Scale across TV, OTT and Connected Screens." *afaqs!*, December 19. https://www.afaqs.com/marketing/as-shark-tank-india-turns-five-sony-bets-on-scale-across-tv-ott-and-connected-screens-10922444 (accessed July 22, 2026).

Jones, Charles I. 2022. "The Past and Future of Economic Growth: A Semi-Endogenous Perspective." *Annual Review of Economics* 14: 125–152.

Kenya Institute of Curriculum Development. 2024. "Junior School Curriculum Design: Pre-Technical Studies, Grade 9." Kenya Institute of Curriculum Development. ISBN 978-9914-43-985-4. https://kicd.ac.ke/cbc-materials/curriculum-designs/grade-nine-designs/ (accessed September 27, 2026).

Khan, Saniya Ahmad. 2024. "Shark Tank India: Impact on India's Growing Startup Mindset." *YourStory*, February 9. https://yourstory.com/2024/02/shark-tank-india-startup-impact (accessed September 26, 2026).

Koning, Rembrand, Sampsa Samila, and John-Paul Ferguson. 2021. "Who Do We Invent For? Patents by Women Focus More on Women's Health, but Few Women Get to Invent." *Science* 372 (6548): 1345–1348.

Laajaj, Rachid, and Karen Macours. 2021. "Measuring Skills in Developing Countries." *Journal of Human Resources* 56 (4): 1254–1295.

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

List, John A. 2024. "Optimally Generate Policy-Based Evidence Before Scaling." *Nature* 626 (7999): 491–499.

Manso, Gustavo. 2011. "Motivating Innovation." *Journal of Finance* 66 (5): 1823–1860.

March, James G. 1991. "Exploration and Exploitation in Organizational Learning." *Organization Science* 2 (1): 71–87.

Munshi, Kaivan. 2019. "Caste and the Indian Economy." *Journal of Economic Literature* 57 (4): 781–834.

Muralidharan, Karthik. 2024. *Accelerating India's Development: A State-Led Roadmap for Effective Governance.* Penguin Viking.

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

Muralidharan, Karthik, and Abhijeet Singh. 2025. "Adapting for Scale: Experimental Evidence on Technology-Aided Instruction in India." NBER Working Paper 34205.

Muralidharan, Karthik, Jishnu Das, Alaka Holla, and Aakash Mohpal. 2017. "The Fiscal Cost of Weak Governance: Evidence from Teacher Absence in India." *Journal of Public Economics* 145: 116–135.

Newell, Allen, and Herbert A. Simon. 1972. *Human Problem Solving.* Englewood Cliffs, NJ: Prentice-Hall.

Nourani, Vesall, Nava Ashraf, and Abhijit Banerjee. 2025. "Learning to Teach by Learning to Learn." Working paper.

OECD. 2024. *PISA 2022 Results (Volume III): Creative Minds, Creative Schools.* Paris: OECD Publishing. https://doi.org/10.1787/765ee8c2-en.

OECD and Eurostat. 2018. *Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation. The Measurement of Scientific, Technological and Innovation Activities*. 4th ed., Paris: OECD Publishing. https://doi.org/10.1787/9789264304604-en.

Press Information Bureau. 2024. "School Innovation Marathon 2024 Attracts Over 1 Lakh Submissions." Press release, NITI Aayog, December 20. https://www.pib.gov.in/PressReleseDetailm.aspx?PRID=2086453 (accessed July 4, 2026).

Press Information Bureau. 2025. "Shri Dharmendra Pradhan Launches Viksit Bharat Buildathon 2025, to Ignite Innovation among School Students across the Country." Press release, Ministry of Education, September 23. https://www.pib.gov.in/PressReleasePage.aspx?PRID=2170192 (accessed August 27, 2026).

Pritchett, Lant, and Amanda Beatty. 2015. "Slow Down, You're Going Too Fast: Matching Curricula to Student Skill Levels." *International Journal of Educational Development* 40: 276–288.

Romer, Paul M. 1990. "Endogenous Technological Change." *Journal of Political Economy* 98 (5, Part 2): S71–S102.

SCERT Delhi. 2021. "Annual Report (2019–2020)." State Council of Educational Research and Training, Government of NCT of Delhi, New Delhi.

Sihag, Garima. 2026. "TGCET Gurukul Result 2026 Out at tgcet.cgg.gov.in: Check 5th, 9th Score Card, Rank." *Careers360*, updated May 15. https://school.careers360.com/articles/tg-gurukul-tgcet-result-2026 (accessed September 26, 2026).

Singh, Abhijeet. 2024. "Improving Administrative Data at Scale: Experimental Evidence on Digital Testing in Indian Schools." *Economic Journal* 134 (661): 2207–2223.

SonyLIV. 2026. "Shark Tank India." https://www.sonyliv.com/shows/shark-tank-india-1700000741 (accessed June 30, 2026).

Telangana State Innovation Cell. 2026. "UNICEF-TSIC School Innovation Challenge." https://teamtsic.telangana.gov.in/unicef-tsic-school-innovation-challenge/ (accessed July 4, 2026).

TGSWREIS. 2026. "About Us." Telangana Social Welfare Residential Educational Institutions Society. https://tgswreis.telangana.gov.in/ (accessed September 26, 2026).

TGTWREIS. 2024. "Academic Results from the Year 2013–2024." Telangana Tribal Welfare Residential Educational Institutions Society. https://tgtwreis.telangana.gov.in/content/results-670e9754c0ecb (accessed July 11, 2026).

TGTWREIS. 2026. "Telangana Tribal Welfare Residential Educational Institutions Society." Official website. https://tgtwreis.telangana.gov.in/ (accessed September 26, 2026).

Toivanen, Otto, and Lotta Väänänen. 2016. "Education and Invention." *Review of Economics and Statistics* 98 (2): 382–396.

TSWREIS. 2022. "TSWREIS Students Excel in SSC Exam." Telangana Social Welfare Residential Educational Institutions Society. https://tgswreis.telangana.gov.in/slider1-SSC-2022.php (accessed July 5, 2026).

Van Reenen, John. 2021. "Innovation and Human Capital Policy." NBER Working Paper 28713.

Weidmann, Ben, and David J. Deming. 2021. "Team Players: How Social Skills Improve Team Performance." *Econometrica* 89 (6): 2637–2657.

World Bank. 2023. "Patent Applications, Residents (per Million People), 2021." *World Development Indicators*, Washington, DC: World Bank. https://data.worldbank.org/indicator/IP.PAT.RESD (accessed July 24, 2026).

World Bank. 2024a. "FLEX 2024: African Ministers and Education Stakeholders Gather in Kigali to Track Progress in Addressing Learning Crisis on the Continent." Press release, November 11. https://www.worldbank.org/en/news/press-release/2024/11/11/flex-2024-afe-ministers-and-education-stakeholders-gather-in-kigali-to-track-progress-in-addressing-learning-crisis.

World Bank. 2024b. *World Development Report 2024: The Middle-Income Trap.* Washington, DC: World Bank.

World Bank. 2026a. "Research and Development Expenditure (% of GDP)." *World Development Indicators*, Washington, DC: World Bank. https://data.worldbank.org/indicator/GB.XPD.RSDV.GD.ZS (accessed August 27, 2026).

World Bank. 2026b. "Researchers in R&D (per Million People), 2020." *World Development Indicators*, Washington, DC: World Bank. https://data.worldbank.org/indicator/SP.POP.SCIE.RD.P6?locations=IN (accessed August 4, 2026).

World Intellectual Property Organization. 2024. "World Intellectual Property Report 2024: Making Innovation Policy Work for Development." World Intellectual Property Organization, Geneva. https://doi.org/10.34667/tind.49284.

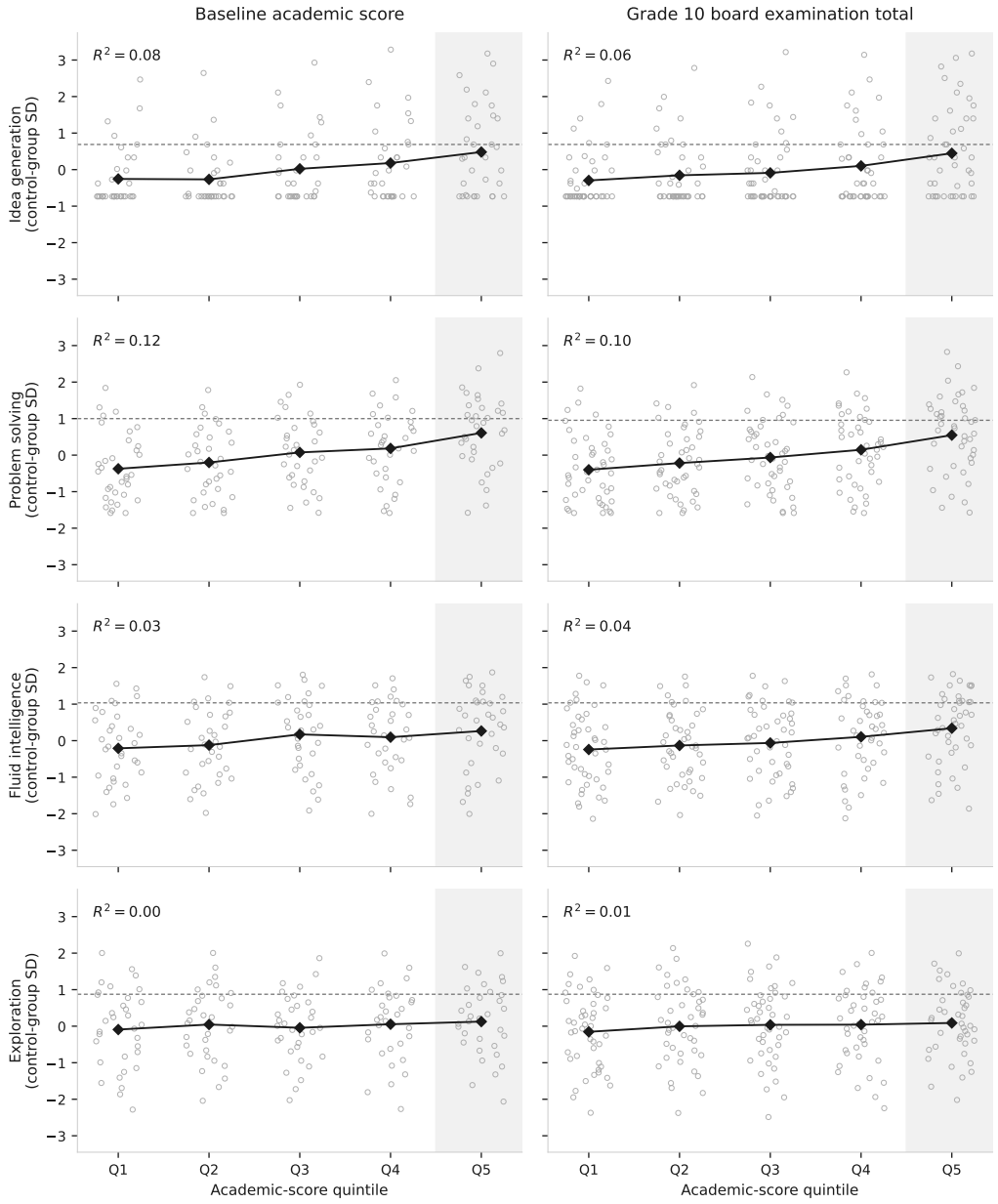


FIGURE 1. ACADEMIC SCORES AND INNOVATION-RELEVANT SKILLS

Notes: Among control students, panels plot four endline skill measures by quintile of baseline academic scores (left; students with an observed baseline score, $N = 1{,}602$) or grade 10 board totals (right; $N = 2{,}123$). Each skill is standardized within the control group. Gray points average approximately ten adjacent skill-ranked students within each academic quintile; diamonds mark quintile means. The shaded region identifies the highest academic quintile, and the dashed line marks the skill's 80th percentile, including ties. Panel R^2 values use continuous academic scores.

Supplemental Appendix

Can Innovation be Taught in Schools?
Experimental Evidence from India

Saloni Gupta

This supplemental appendix contains additional results, measurement details, and study materials accompanying the main paper.

A. Additional Tables and Figures

B. The Curriculum: Content, Delivery, and Implementation Fidelity

C. The Innovation Challenge: Measurement and Materials

D. Skill Tasks: Design, Scoring, and Process Evidence

E. Academic Outcomes

F. Student Feedback at Midline

G. Deviations from the Pre-Analysis Plan

A. Additional Tables and Figures

TABLE A.1—SAMPLE AND INNOVATION CHALLENGE FUNNEL

	Control	Treatment	Total
Panel A. Individual sample, by wave			
Enrolled grade 8 cohort	2,807	2,835	5,642
Baseline assessment	2,597	2,618	5,215
Midline assessment	2,729	2,634	5,363
Endline assessment	2,126	2,262	4,388
Grade 10 board records	2,801	2,832	5,633
Panel B. Endline Innovation Challenge (team analysis)			
Students			
Assessed at endline	2,126	2,262	4,388
On a submitted team	1,558	1,828	3,386
Non-submitters (not on a submitted team)	568	434	1,002
Teams			
Non-submitter teams (grouped ≤ 4 within school $\times$ arm)	155	122	277
Idea-rating estimation sample	723	770	1,493

Notes: Panel A traces the study cohort across data sources by treatment arm. The enrolled grade 8 cohort is the set of students enrolled at baseline; the baseline, midline, and endline rows count students assessed in each wave; the grade 10 row counts students with a board examination result (pass or fail). Grade 10 records cover 99.8% of the enrolled cohort, and 77% of the baseline-assessed roster is matched to a grade 10 record. The wave-attendance comparisons reported in Section III use school-level regressions of students assessed divided by the enrolled cohort on treatment and tribal $\times$ girls $\times$ rural stratum fixed effects, with heteroskedasticity-robust (HC1) standard errors. Panel B describes the endline Innovation Challenge team analysis. Students assessed at endline are split into those linked to a submitted team and non-submitters; non-submitters are grouped into teams of at most four within school-by-arm cells. The idea-rating estimation sample combines the scored submitted teams that retain at least one endline member with the non-submitter teams. Team rosters name 5,049 distinct students, 661 of whom were not individually assessed at endline.

TABLE A.2—SCHOOL CHARACTERISTICS AND BALANCE

	Control mean	Treatment mean	Treatment− Control	*p*-value	*N*
Rented building	0.250	0.375	0.125	0.233	80
Has library	0.500	0.500	0.000	1.000	80
Has playground	0.125	0.150	0.025	0.749	80
English medium	0.825	0.850	0.025	0.765	80
Has ramp	0.725	0.675	-0.050	0.631	80
Has computer	0.675	0.650	-0.025	0.816	80
Boys' toilets	6.175	6.800	0.625	0.806	80
Girls' toilets	9.250	8.300	-0.950	0.758	80
Degraded boundary wall	0.075	0.075	0.000	1.000	80
No boundary wall	0.175	0.175	0.000	1.000	80

Notes: The unit of observation is the school ($N = 80$). Each row reports means by treatment arm and the treatment-control difference from a school-level regression of the characteristic on treatment assignment, with heteroskedasticity-robust standard errors. Toilet rows are counts; all other characteristics are indicators.

TABLE A.3—SAMPLE COVERAGE AND ATTRITION

	Control mean	Treatment mean	Treatment − Control	*p*-value	*N*
Panel A. Observation rates among baseline-assessed students (%)					
Observed in midline assessment	69.6	67.2	-2.2	0.231	5,215
Observed in endline assessment	61.7	63.9	2.4	0.292	5,215
Linked to grade 10 records	77.7	76.8	-0.7	0.677	5,215
Panel B. Differential attrition: treatment × baseline covariate on endline observation					
Treatment × Age			2.06	0.122	5,215
Treatment × Mathematics			-0.30	0.850	5,215
Treatment × Science			-1.44	0.397	5,215
Treatment × Fluid intelligence			0.40	0.791	5,215
Treatment × Emotional intelligence			0.11	0.951	5,215
Treatment × Big Five			1.82	0.197	5,215
Treatment × Risk preference			0.61	0.644	5,215
Joint test *p*-value				0.501	5,215
Attriter difference: Grade 10 mathematics (ext., SD)			0.036	0.580	5,632
Attriter difference: Grade 10 science (ext., SD)			-0.026	0.572	5,632
Panel C. Baseline balance among endline-observed students					
Age (years)	13.261	13.246	-0.037	0.421	3,276
Household SES index (SD)	0.061	-0.056	-0.130	0.156	2,884
Household computer literacy (0–1)	0.376	0.413	0.045	0.121	3,276
Mathematics score (%)	42.8	42.2	-0.1	0.953	3,276
Science score (%)	53.4	53.3	0.3	0.801	3,276
Fluid intelligence (% correct)	59.5	59.8	0.3	0.829	3,276
Emotional intelligence (% correct)	53.4	52.8	-0.3	0.737	3,276
Big Five index (SD)	0.040	-0.001	-0.026	0.524	3,276
Risk preference (% of endowment)	62.2	63.0	0.6	0.534	3,272
Joint test *p*-value				0.804	3,276

Notes: The sample is students who completed the baseline assessment. Panel A reports the share observed in each follow-up record by treatment arm, in percent; the treatment-control difference is the coefficient on treatment from a regression of the observation indicator on treatment and randomization strata, with standard errors clustered at the school level. Panel B reports coefficients on treatment interacted with standardized baseline covariates in a regression of endline observation on treatment, the covariates, their interactions, and strata (coefficients in percentage points per standard deviation); the joint test is a Wald test of all interactions, and missing covariate values are imputed at sample means in the interaction regressions and joint tests. The final two Panel B rows compare students not observed at endline with those observed, on externally graded grade 10 board examination scores (in standard deviations), conditional on strata. Panel C reports baseline covariate means by arm among students observed in the endline assessment, with the same specification as the baseline balance table. Variable definitions and standard deviations follow Table 2.

TABLE A.4—SUBMITTING-TEAM CHARACTERISTICS AND BOUNDS ON CONDITIONAL RATINGS

	Control mean	Treatment mean	Treatment−Control	*p*-value	*N*
Panel A. Team-average baseline characteristics					
Members with baseline records (%)	73.7	69.6	-3.4	0.136	1,225
Age (years)	13.294	13.274	-0.042	0.337	1,225
Academic score (SD)	0.012	-0.061	-0.048	0.577	1,225
Fluid intelligence (SD)	0.001	-0.039	-0.034	0.581	1,225
Emotional intelligence (SD)	0.035	-0.051	-0.070	0.285	1,225
Exploration index (SD)	0.014	-0.009	-0.019	0.626	1,225
Risk preference (SD)	-0.029	0.015	0.032	0.461	1,225
Household SES index (SD)	0.044	-0.056	-0.114	0.210	1,225
Household computer literacy (0–1)	0.374	0.394	0.026	0.367	1,225
Joint test *p*-value				0.789	1,225
Team-formation margins					
Team size (members)	3.83	3.90	0.08	0.130	1,225
Within-team SD: academic score	0.684	0.655	-0.025	0.415	1,221
Within-team SD: fluid intelligence	0.774	0.719	-0.051	0.051	1,221
Within-team SD: exploration index	0.483	0.486	0.006	0.771	1,221
Joint test *p*-value				0.061	1,221
Panel B. Lee (2009) bounds on conditional quality differences					
	T−C	Lower	Upper		*N*
Novelty	0.226	0.150	0.380		1,225
Value-Add	0.162	0.082	0.318		1,225
Feasibility	0.080	-0.018	0.218		1,225
Willingness to Fund	0.158	0.047	0.292		1,225

Notes: The sample is the 1,225 scored team submissions to the Innovation Challenge; 1,216 of these have an endline-assessed member and enter Table 4. Panel A reports team-average baseline characteristics by treatment arm: averages are taken over team members, with missing member values imputed as in the estimation samples (Section V), and the first row reports the share of members with observed baseline records. The treatment−control difference is the coefficient on treatment from a regression of the characteristic on treatment and randomization strata, with standard errors clustered at the school level; the joint test regresses treatment on all eight characteristics and strata and tests that the characteristics are jointly unrelated to treatment. The team-formation margins report team size and the within-team standard deviation of members' baseline scores, computed over members with observed baseline records for teams with at least two such members; their joint test is defined analogously over the four margins. Panel B reports Lee (2009) bounds on the treatment−control difference in idea quality among submitted teams: 84.2 percent of treatment and 78.7 percent of control teams in the endline cohort submitted ideas, so each bound trims 6.6 percent of the treatment teams from the top (lower bound) or the bottom (upper bound) of the outcome distribution. Under random assignment and monotonic selection—treatment does not prevent a team from submitting—these bounds address selection into submission. Panel B's T−C column and bounds are unadjusted comparisons in submitted-control-group standard deviations, whereas Panel A's differences adjust for strata. Bound endpoints are point estimates rather than confidence limits. Outcomes are defined in Table 4.

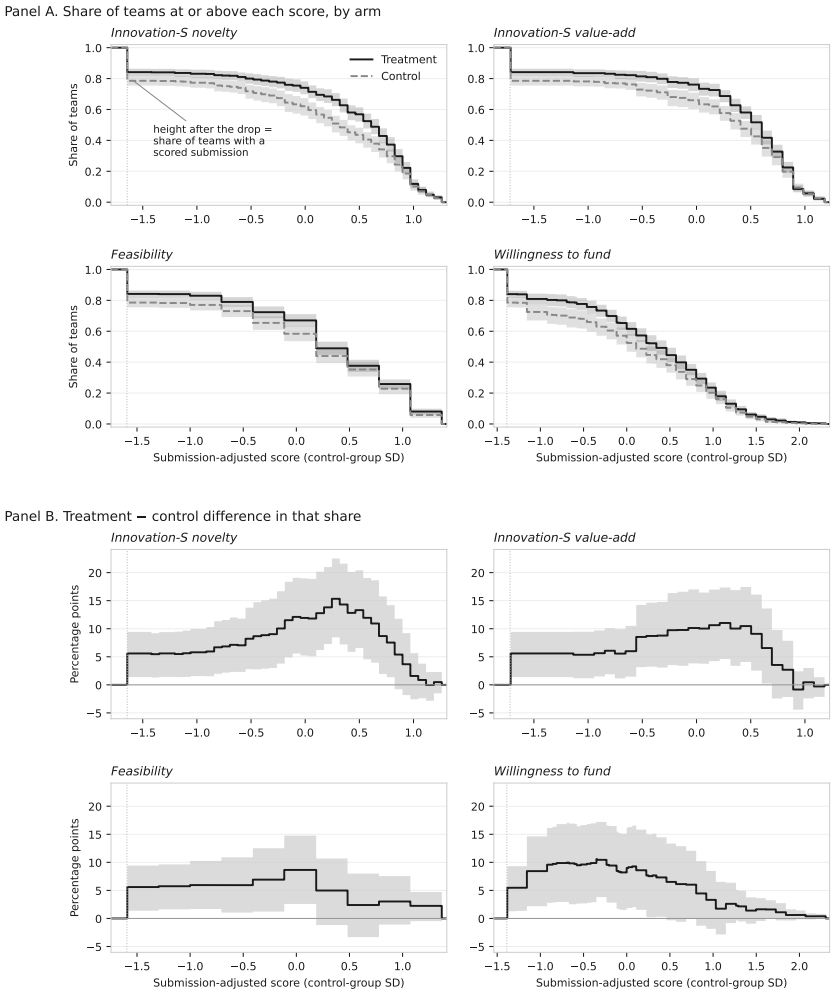


FIGURE A.1. IDEA OUTPUT ACROSS THE DISTRIBUTION OF RATED QUALITY

Notes: The sample is the team universe of Table 4, rows 2–5: 1,493 teams, the 1,216 scored submitted teams with at least one endline-assessed member together with 277 latent teams of up to four members formed within school-by-arm cells from the 1,002 endline students not linked to a scored submission. Latent teams carry the value of producing no idea, and scores are standardized to the control group of this sample, so the horizontal axis is the submission-adjusted scale of Table 4. Every latent team carries the same no-idea value and the number of latent teams in each school does not depend on how students are grouped, so the curves are identical across the 1,000 random groupings used for estimation. Panel A plots, for each rating and arm, the share of teams at or above each score. The dotted vertical line marks the no-idea value: each curve drops there by the share of teams without a scored submission, so the vertical gap between arms just right of the line is the participation margin, and the remaining gap shows differences between arms in the percentage of teams meeting each quality threshold. Panel B plots the treatment−control difference in the same share, in percentage points. Shaded regions are pointwise 95 percent confidence bands from a school-cluster bootstrap that resamples schools with replacement within arm (999 draws); treatment was assigned at the school level. The corresponding estimates and standard errors are in Table 4; conditional estimates among submitted teams are in Appendix Table A.5.

TABLE A.5—CONDITIONAL IDEA RATINGS AMONG SUBMITTED INNOVATION CHALLENGE TEAMS

Outcome	Treatment effect	SE	*N*
Innovation-S novelty subscale (SD)	0.265***	(0.070)	1,225
Innovation-S value-add subscale (SD)	0.202***	(0.069)	1,225
Feasibility rating (SD)	0.100	(0.062)	1,225
Evaluator willingness to fund (SD)	0.206***	(0.073)	1,225

Notes: This appendix table reports conditional treatment effects among Innovation Challenge teams that submitted a scored idea. Outcomes use the same evaluator ratings as Table 4: the Innovation-S novelty and value-add subscales, feasibility rating, and evaluator willingness to fund. Ratings are standardized using the submitted control group, so coefficients are in control-group standard deviations. Submission is a post-treatment outcome. All regressions control for randomization strata, school infrastructure, team-average baseline covariates (members' baseline academic achievement, exploration index, fluid intelligence, emotional intelligence, household SES index, computer literacy, and age), team size, the number of members with baseline records, and the number of members not matched to a student record. Standard errors are clustered by school. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE A.6—JURY FUNDING RAISED AT THE SHARK TANK FINALS

Outcome	Control mean	Treatment effect	% of control	*p*-value	*N*
Total jury funding raised (no-shows = 0)	Rs 6,300	+Rs 1,709	+27%	0.28	80
with jury-panel fixed effects (attendees)	Rs 6,632	+Rs 1,064	+16%	0.40	78

Notes: Among Innovation Challenge finalists (the top teams from each arm invited to the Shark Tank-style event), the outcome is the total funding a team raised, summed across its jurors, in rupees; the jurors' allocations were disbursed to the teams, in contrast to the evaluators' hypothetical willingness-to-fund measure. The two control teams that were invited and travel-funded but did not attend are coded as raising zero. Row 1's *p*-value is from school-level randomization inference (treatment is randomized at the school level; permuted within strata, 5,000 draws); it is computed on the 78 attended teams, whose schools are observed. The second row adds jury-panel fixed effects among attendees as a robustness check, with its *p*-value from school-clustered standard errors; one jury saw only control teams, so the fixed-effects row identifies the effect from the remaining juries. The attendee sample spans approximately 29 schools. See Figure 3 for the funding distributions.

TABLE A.7—PROBLEM SOLVING: ROBUSTNESS TO THE DIFFICULTY CHOICE

Outcome coding	Treatment effect	SE	p	N
Difficulty-weighted score (Table 5 measure)	0.142**	(0.064)	0.029	4,388
Score residualized on the difficulty choice	0.140**	(0.063)	0.029	4,388
Unweighted score (first principal component)	0.136**	(0.064)	0.038	4,388
Hard-puzzle rounds only	0.113**	(0.055)	0.043	3,496

Notes: The table re-estimates the Table 5 Panel C problem-solving effect under alternative treatments of the easy-versus-hard difficulty choice, which the main measure builds into the score by weighting hard puzzles 1.5 times. Row 1 repeats the main measure, the difficulty-weighted average score across the five Sudoku rounds. Row 2 residualizes the weighted score on the number of hard puzzles the student faced (including the imposed hard puzzle in round 1 for the randomly selected half of students) before standardizing, so the choice margin is removed from the score. Row 3 scores performance with no difficulty adjustment, as the first principal component of the five raw round scores. Row 4 averages scores over only the rounds in which the student faced the hard puzzle; the sample is the 3,496 students who faced at least one. All outcomes are standardized to the control group. The choice margin itself is the challenge-seeking outcome of Table 5 Panel D. Every regression controls for randomization strata, school infrastructure, individual covariates (age, socioeconomic index, computer skills), and baseline fluid intelligence with a missing-value indicator, matching the Table 5 Panel C specification. Standard errors, in parentheses, are clustered at the school level, the level of treatment assignment; p-values use a t-distribution with degrees of freedom equal to the number of schools minus one. The rows are alternative codings of one outcome rather than separate hypotheses, so no false-discovery-rate adjustment applies. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

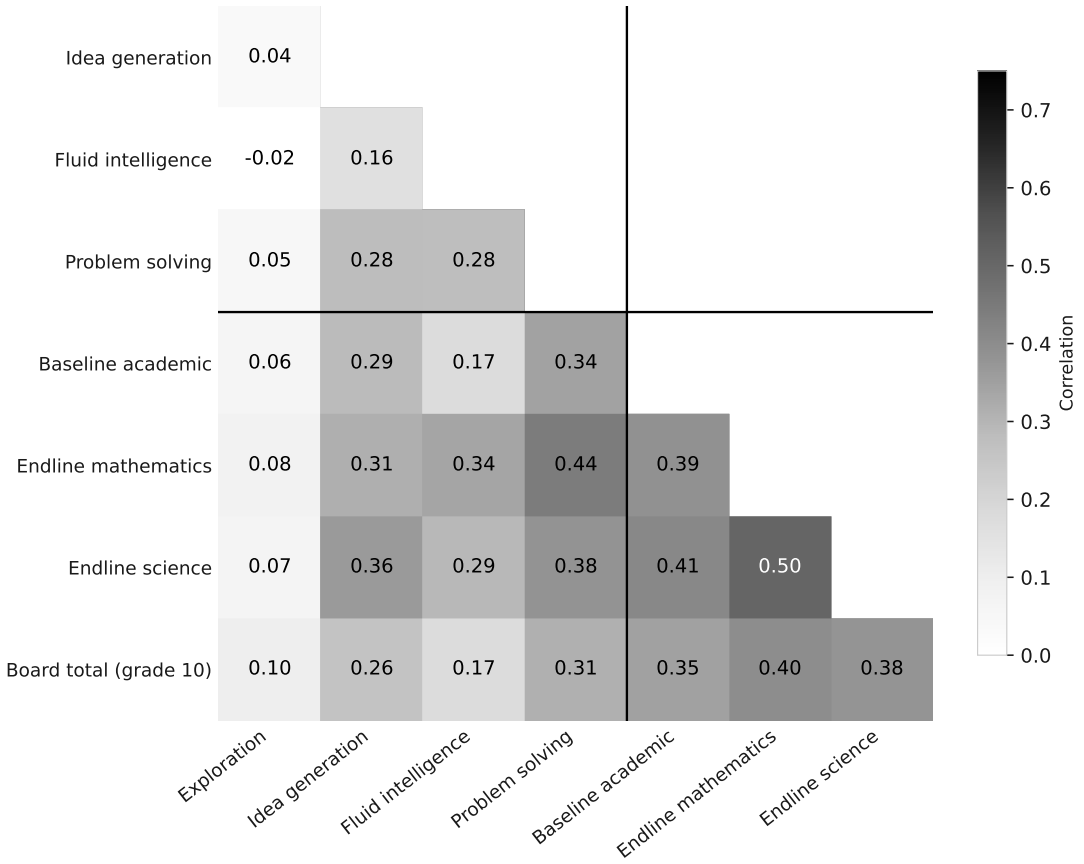


FIGURE A.2. CORRELATIONS AMONG SKILL AND ACADEMIC MEASURES

Notes: Pairwise correlations among the four endline skill measures and four academic measures, computed over the 1,600 control students with a non-imputed baseline academic score and all eight measures observed. Each correlation appears once; darker cells indicate stronger correlations, and the solid lines separate the skill block from the academic block.

TABLE A.8—HETEROGENEOUS TREATMENT EFFECTS: TRIBAL-WELFARE MANAGEMENT AND LOCATION

	By welfare-system management			By location			
	Tribal (ST)	Social (SC)	p	Urban	Rural	p	N
Panel A. Innovation Challenge production (Table 4 outcomes)							
Participation (pp)	11.9***	4.6	0.109	3.6	10.5***	0.161	4,388
	(3.3)	(2.8)		(4.1)	(2.4)		
Novelty (SD)	0.274***	0.109	0.143	0.038	0.273***	0.050	1,493
	(0.076)	(0.078)		(0.093)	(0.068)		
Value-add (SD)	0.228***	0.072	0.152	-0.022	0.242***	0.021	1,493
	(0.077)	(0.073)		(0.087)	(0.067)		
Feasibility (SD)	0.185***	0.041	0.138	-0.034	0.191***	0.026	1,493
	(0.068)	(0.065)		(0.073)	(0.062)		
Willingness to fund (SD)	0.287***	0.091	0.105	0.076	0.246***	0.219	1,493
	(0.091)	(0.080)		(0.102)	(0.083)		
Panel B. Innovation-relevant skills (Table 5 outcomes)							
Exploration index (points)	0.092**	0.095***	0.961	0.091**	0.095***	0.942	4,388
	(0.040)	(0.030)		(0.037)	(0.033)		
Idea generation (SD)	0.149	0.460***	0.057	0.349**	0.294***	0.766	4,388
	(0.103)	(0.118)		(0.159)	(0.086)		
Idea evaluation (SD)	-0.034	0.220**	0.051	0.067	0.119	0.690	4,388
	(0.088)	(0.086)		(0.108)	(0.072)		
Written expression (SD)	-0.115	0.178*	0.032	0.032	0.045	0.934	4,388
	(0.096)	(0.089)		(0.125)	(0.082)		
Fluid intelligence (SD)	-0.004	0.213***	0.050	0.104	0.115*	0.934	4,388
	(0.073)	(0.074)		(0.093)	(0.067)		
Problem solving (SD)	0.088	0.189**	0.469	0.017	0.210***	0.185	4,388
	(0.095)	(0.094)		(0.120)	(0.076)		
Emotional intelligence (SD)	-0.051	0.103	0.108	0.070	0.009	0.604	4,388
	(0.066)	(0.068)		(0.094)	(0.061)		
Challenge-seeking (SD)	-0.015	0.049	0.685	0.041	0.007	0.831	4,388
	(0.111)	(0.110)		(0.131)	(0.094)		
Panel C. Academic outcomes (Table 6 outcomes)							
Endline mathematics (SD)	-0.175**	0.142	0.019	0.142	-0.089	0.184	4,388
	(0.087)	(0.097)		(0.131)	(0.088)		
Endline science (SD)	-0.045	0.090	0.261	0.037	0.021	0.910	4,388
	(0.075)	(0.093)		(0.107)	(0.080)		
Grade 10 total score (SD)	-0.011	0.266	0.239	0.101	0.156	0.827	5,633
	(0.157)	(0.168)		(0.200)	(0.142)		

Notes: Specification, samples, and inference follow Table 7 exactly; the moderating characteristic is tribal-welfare management (reference: tribal-welfare, ST) in the first block and rural location (reference: urban) in the second, each removed from the strata controls and entered directly with its interaction. Standard errors, in parentheses, are clustered at the school level; stars refer to each group effect's clustered p-value. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE A.9—GENDER AND BASELINE-SCORE INTERACTIONS: JOINT AND WITHIN-SUBSAMPLE ESTIMATES

Interaction	Coefficient	SE	p	N
Idea generation: $T\times$ Girls' school	0.337**	(0.142)	0.020	4,388
Idea generation: $T\times$ Above-median score	0.198**	(0.099)	0.049	4,388
Fluid intelligence: $T\times$ Girls' school	0.048	(0.092)	0.605	4,388
Fluid intelligence: $T\times$ Above-median score	0.194**	(0.081)	0.018	4,388
Idea generation: $T\times$ Girls' school, SC only	0.619***	(0.201)	0.004	2,298
Idea generation: $T\times$ Girls' school, ST only	0.050	(0.139)	0.722	2,090

Notes: The first four rows jointly estimate the two interactions in one regression per outcome, $Y=\beta_0+\beta_1 T+\beta_2 G+\beta_3(T\times G)+\beta_4 A+\beta_5(T\times A)+\mathbf{X}\gamma+\varepsilon$, where G indicates a girls-only school and A indicates an above-median baseline academic score. The last two rows estimate the Table 7 gender interaction separately within the 40 social-welfare (SC) and the 40 tribal-welfare (ST) schools; gender and welfare system are fully crossed in the design, 20 schools per cell. All specifications otherwise follow Table 7, including each outcome's baseline control; standard errors are clustered at the school level, and p-values use the clustered t-distribution. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

TABLE A.10—ALTERNATIVE SPECIFICATIONS: INNOVATION CHALLENGE PRODUCTION

Outcome	(1) No controls	(2) Strata indicators	(3) Strata + baseline	(4) Main specification	(5) Stratum-cell FE	N
Participation (pp)	7.5**	7.7***	7.8***	8.0***	8.0***	4,388
	(2.9)	(2.4)	(2.3)	(2.1)	(2.1)	
Novelty (SD)	0.229***	0.238***	0.242***	0.187***	0.187***	1,493
	(0.069)	(0.058)	(0.057)	(0.054)	(0.054)	
Value-add (SD)	0.190***	0.199***	0.200***	0.146***	0.146***	1,493
	(0.066)	(0.055)	(0.055)	(0.052)	(0.052)	
Feasibility (SD)	0.152**	0.161***	0.157***	0.109**	0.109**	1,493
	(0.062)	(0.049)	(0.050)	(0.047)	(0.046)	
Willingness to fund (SD)	0.205***	0.214***	0.219***	0.184***	0.183***	1,493
	(0.077)	(0.069)	(0.067)	(0.062)	(0.061)	
Strata indicators	No	Yes	Yes	Yes	No	
Saturated stratum-cell fixed effects	No	No	No	No	Yes	
Baseline controls + missing flags	No	No	Yes	Yes	Yes	
School infrastructure controls	No	No	No	Yes	Yes	
Individual or team composition controls	No	No	No	Yes	Yes	

Notes: The table re-estimates every Table 4 row under five specifications. Column (1) regresses the outcome on treatment alone; column (2) adds the three randomization-strata indicators (tribal-welfare management, girls-only composition, rural location); column (3) adds the row's baseline controls with their missing-value flags: baseline mathematics and science scores for the participation row, and the seven team baseline composites (means over matched members) for the rating rows; column (4) repeats the main specification of Table 4 and reproduces it exactly; column (5) keeps the main specification but replaces the three strata indicators with the eight saturated stratum-cell fixed effects formed by their interaction. Participation is in percentage points on the endline individual sample. The rating rows keep the submission-adjusted design in every column: endline students not linked to a submitted scored team are randomly partitioned within school-by-arm cells into latent teams of four assigned the mean non-submitter rating, the reported coefficient is the average treatment effect across 1,000 partitions, and its variance is the average school-clustered variance across partitions plus the between-partition variance of the treatment effect. Sample sizes are identical across the five columns within each row. Standard errors, in parentheses, are clustered at the school level, the level of treatment assignment; significance stars refer to the clustered p-value from a t-distribution with degrees of freedom equal to the number of schools minus one. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE A.11—ALTERNATIVE SPECIFICATIONS: INNOVATION-RELEVANT SKILLS

Outcome	(1) No controls	(2) Strata indicators	(3) Strata + baseline	(4) Main specification	(5) Stratum-cell FE	N
Panel A. Search behavior in the exploration task						
Exploration index (points)	0.088***	0.089***	0.092***	0.094***	0.096***	4,388
	(0.030)	(0.026)	(0.026)	(0.025)	(0.024)	
Initiation (pp)	5.6***	5.7***	5.7***	5.6***	5.7***	4,388
	(2.0)	(1.9)	(1.9)	(1.8)	(1.8)	
Breadth	0.138	0.144	0.147	0.161*	0.163*	4,388
	(0.105)	(0.094)	(0.094)	(0.092)	(0.088)	
Discovery	0.090**	0.091**	0.095**	0.101**	0.104***	4,388
	(0.044)	(0.042)	(0.041)	(0.039)	(0.038)	
Convergence	-0.252*	-0.255**	-0.265**	-0.256**	-0.262**	4,388
	(0.132)	(0.128)	(0.123)	(0.115)	(0.112)	
Panel B. Creativity						
Idea generation (SD)	0.310***	0.313***	0.316***	0.314***	0.314***	4,388
	(0.096)	(0.084)	(0.083)	(0.078)	(0.076)	
Idea evaluation (SD)	0.099	0.103	0.106	0.100*	0.099*	4,388
	(0.088)	(0.066)	(0.065)	(0.060)	(0.058)	
Written expression (SD)	0.031	0.036	0.039	0.040	0.041	4,388
	(0.095)	(0.075)	(0.074)	(0.065)	(0.063)	
Panel C. Analytical reasoning						
Fluid intelligence (SD)	0.108	0.110	0.113*	0.111**	0.109**	4,388
	(0.074)	(0.069)	(0.057)	(0.050)	(0.049)	
Problem solving (SD)	0.102	0.107	0.109	0.142**	0.141**	4,388
	(0.096)	(0.082)	(0.074)	(0.064)	(0.063)	
Panel D. Emotional intelligence and challenge-seeking						
Emotional intelligence (SD)	0.006	0.012	0.021	0.031	0.030	4,388
	(0.091)	(0.062)	(0.051)	(0.048)	(0.046)	
Challenge-seeking (SD)	0.009	0.009	0.015	0.019	0.022	4,388
	(0.083)	(0.081)	(0.081)	(0.078)	(0.072)	
Strata indicators	No	Yes	Yes	Yes	No	
Saturated stratum-cell fixed effects	No	No	No	No	Yes	
Baseline control + missing flag	No	No	Yes	Yes	Yes	
School infrastructure controls	No	No	No	Yes	Yes	
Individual controls	No	No	No	Yes	Yes	

Notes: The table re-estimates every Table 5 row under five specifications; outcome definitions, samples, and panel baselines are as there. Column (1) regresses the outcome on treatment alone; column (2) adds the three randomization-strata indicators (tribal-welfare management, girls-only composition, rural location); column (3) adds the panel's baseline control with its missing-value flag (the baseline exploration index for Panels A and B, baseline fluid intelligence for Panel C, and, in Panel D, the baseline eyes test for emotional intelligence and baseline conscientiousness for challenge-seeking); column (4) repeats the main specification of Table 5, adding school infrastructure and individual controls (age, household assets, computer skills), and reproduces it exactly; column (5) keeps the main specification but replaces the three strata indicators with the eight saturated stratum-cell fixed effects formed by their interaction. Initiation is reported in percentage points; Breadth, Discovery, and Convergence are in natural units. The exploration index is the equal-weighted mean of the four control-standardized components, with Convergence reverse-signed. Outcomes in Panels B–D are standardized to have control-group mean zero and standard deviation one. Sample sizes are identical across the five columns within each row. Standard errors, in parentheses, are clustered at the school level, the level of treatment assignment; significance stars refer to the clustered p-value from a t-distribution with degrees of freedom equal to the number of schools minus one. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE A.12—ALTERNATIVE SPECIFICATIONS: ACADEMIC OUTCOMES

Outcome	(1) No controls	(2) Strata indicators	(3) Strata + baseline	(4) Main specification	(5) Stratum-cell FE	N
Panel A. Endline assessments						
Mathematics	-0.044	-0.039	-0.022	-0.007	-0.010	4,388
	(0.110)	(0.095)	(0.078)	(0.067)	(0.063)	
Science	0.008	0.012	0.018	0.027	0.026	4,388
	(0.095)	(0.082)	(0.066)	(0.061)	(0.058)	
Panel B. Grade 10 board examinations						
Mathematics	0.104	0.106	0.112	0.112	0.107	5,632
	(0.157)	(0.155)	(0.156)	(0.147)	(0.143)	
Science	-0.013	-0.010	-0.008	-0.006	-0.012	5,632
	(0.133)	(0.130)	(0.131)	(0.119)	(0.113)	
Social studies	0.001	0.004	0.010	0.023	0.015	5,632
	(0.109)	(0.102)	(0.104)	(0.093)	(0.089)	
First language	0.115	0.119	0.130	0.149*	0.145*	5,632
	(0.106)	(0.094)	(0.094)	(0.088)	(0.086)	
Second language	0.172	0.175	0.182	0.199*	0.196*	5,631
	(0.111)	(0.108)	(0.110)	(0.106)	(0.104)	
Third language	0.126	0.126	0.135	0.153*	0.153*	5,631
	(0.098)	(0.089)	(0.092)	(0.086)	(0.084)	
Total score	0.112	0.116	0.126	0.135	0.130	5,633
	(0.125)	(0.120)	(0.123)	(0.115)	(0.111)	
Passed (pp)	0.53	0.54	0.59	0.64	0.61	5,642
	(0.91)	(0.90)	(0.89)	(0.83)	(0.77)	
Strata indicators	No	Yes	Yes	Yes	No	
Saturated stratum-cell fixed effects	No	No	No	No	Yes	
Baseline controls + missing flags	No	No	Yes	Yes	Yes	
School infrastructure controls	No	No	No	Yes	Yes	
Individual controls	No	No	No	Yes	Yes	

Notes: The table re-estimates the Table 6 academic rows under five specifications; outcome definitions and samples are as there, and the behavioral rows (homework share, attendance) are not part of the specification ladder. Column (1) regresses the outcome on treatment alone; column (2) adds the three randomization-strata indicators (tribal-welfare management, girls-only composition, rural location); column (3) adds the row's baseline control(s) with missing-value flags (baseline mathematics for mathematics, baseline science for science, and both for the remaining grade 10 rows); column (4) repeats the main specification of Table 6, adding school infrastructure and individual controls (age, household assets, computer skills), and reproduces it exactly; column (5) keeps the main specification but replaces the three strata indicators with the eight saturated stratum-cell fixed effects formed by their interaction. Test-score outcomes are standardized to the control group; Passed is in percentage points with absentee and withheld results coded as not passing. Sample sizes are identical across the five columns within each row. Standard errors, in parentheses, are clustered at the school level, the level of treatment assignment; significance stars refer to the clustered p-value from a t-distribution with degrees of freedom equal to the number of schools minus one. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

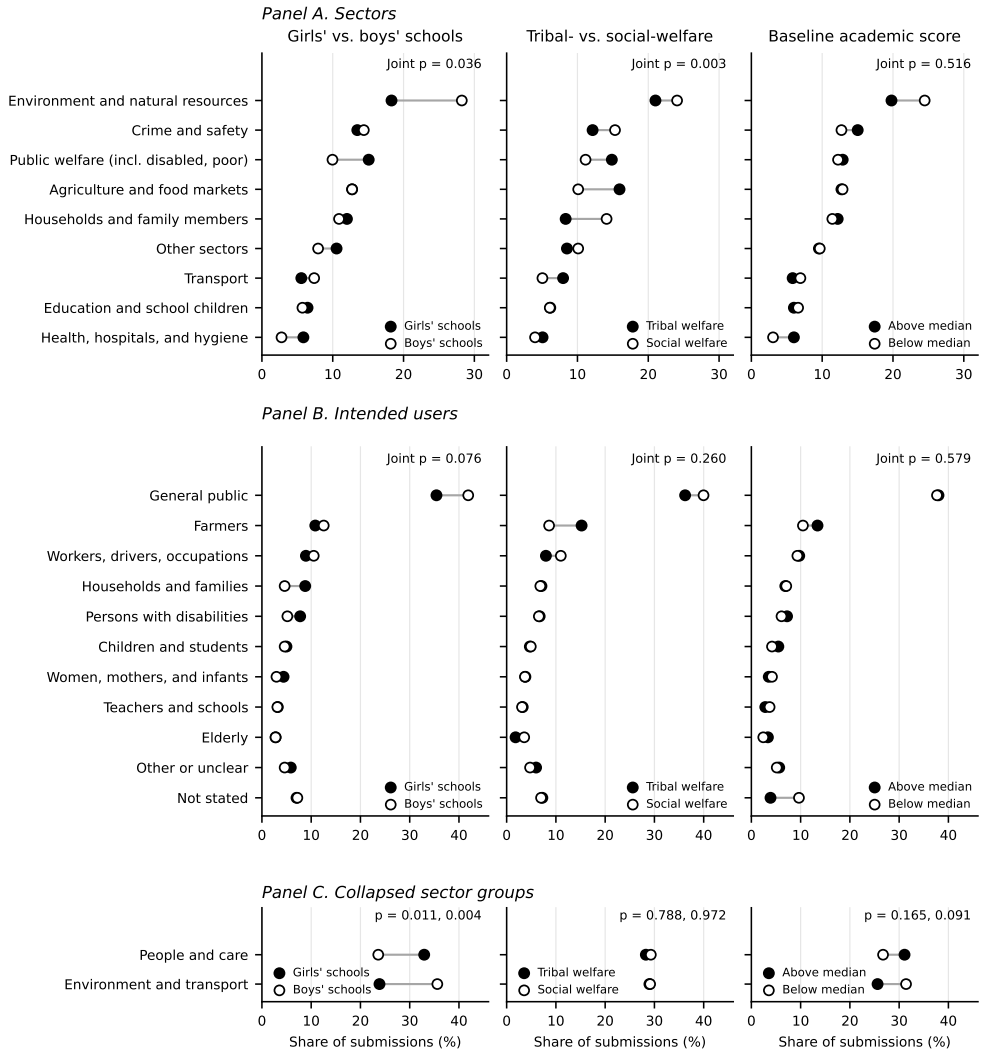


FIGURE A.3. DIRECTION OF IDEAS: SECTORS AND INTENDED USERS BY TEAM ATTRIBUTES

Notes: The figure is descriptive: it reports what the 1,225 scored Innovation Challenge submissions set out to solve (Panel A, the nine-category sector the team selected on the submission form) and for whom (Panel B, intended users classified from the submission's short free-text field with the deterministic keyword rules in the replication code; unmatched descriptions, 5.3 percent, appear as Other or unclear, and blank fields as Not stated), pooling both arms. Within each panel, filled and open circles are the two groups' shares of submissions, the gray segment is their gap, and categories are ordered by the pooled share. Schools are single-sex, so the first column compares girls' and boys' schools (683 and 542 submissions); the second compares tribal-welfare and social-welfare schools (552 and 673); the third splits teams at the median baseline academic score of control-group submitted teams, computed over matched members (566 above, 621 below; 38 teams with no matched baseline member are excluded from this column). Because teams form after treatment and submission is voluntary, the sample comprises submitted ideas rather than a random sample of students, and no causal interpretation is attached. Panels A and B report a joint Wald test that the category shares are equal across the two groups, using school-clustered variance from a stacked linear probability system and referred to an F distribution with the number of categories minus one and 79 degrees of freedom. Panel C collapses the nine sectors into two groups, people and care (public welfare; health, hospitals, and hygiene; households and family members) and environment and transport, and reports the two one-degree tests, defined analogously, in row order. The collapsed sector groups were defined after inspecting Panel A. The test results, with an Anderson sharpened false-discovery-rate adjustment across the eight group tests, are reported in Section VI.

B. The Curriculum: Content, Delivery, and Implementation Fidelity

TABLE B.1—THE THINK & MAKE CURRICULUM AND DELIVERY MODEL

Study year and block	Instructional progression	What students do	Required work product
Panel A: Curriculum Sequence and Student Work			
Year 1, Block 1: Planet	Guided application of the problem-solving cycle to environmental problems	Observe resource use; identify and select a problem; interview affected people; investigate causes and effects; generate and evaluate solutions; build, test, and present a prototype	Problem inventory; interview notes; cause-and-effect analysis; idea-selection rubric; prototype; presentation
Year 1, Block 2: People	Guided application of the same cycle to problems affecting people's quality of life	Identify problems through observation and interviews; investigate a selected problem; generate alternatives; choose a solution; prototype, test, and present	Problem statement; interview evidence; written problem story; idea-selection rubric; prototype; presentation
Year 2, Block 1: Design Evolutions	Worked examples, followed by constrained redesign and independent problem solving	Study how everyday objects evolved; solve a guided story-card problem; redesign an assigned object for a specified user; identify and score problems and ideas from students' own surroundings; prototype and present	Story-card solution; mock-up and sketch prototypes; problem- and idea-ranking sheets; resource plan; final prototype and presentation
Year 2, Block 2: Smarter Everyday	Technical experimentation, followed by guided application and independent technology integration	Build input–sensor–output circuits; experiment with sensors and outputs; evaluate when sensor use is useful; solve guided sensor problems; generate and score independent applications; prototype and present	Circuit exercises; sensor-use critiques; sensor-application ideas; storyboard, sketch, or working mock-up; resource plan; final presentation

Continued on next page

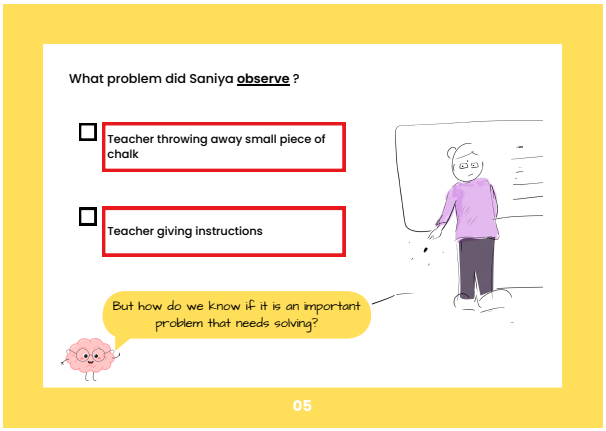
Table B.1 continued

Panel B: Delivery Model

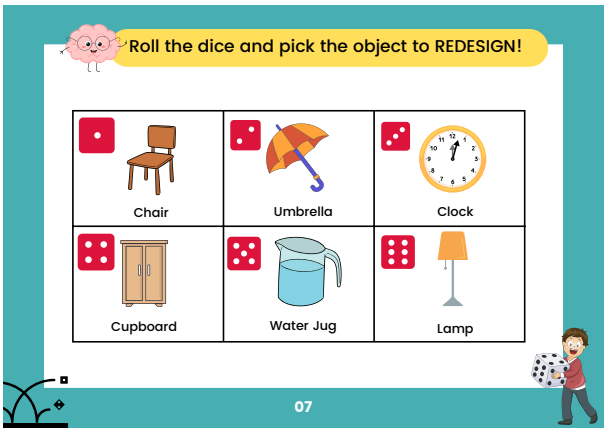
Student organization	Students worked primarily in fixed teams of four. Each team was paired with a co-team for structured feedback.
Student leaders	Two trained student leaders facilitated each classroom, divided facilitation and observation responsibilities, managed time and materials, and guided teams through the workbook instructions.
Teacher role	A teacher was expected to be present to support classroom management, mediate problems, advise teams when needed, and help organize showcases, but did not deliver the curriculum.
Curriculum materials	Step-by-step student workbooks, story and activity cards, presentation cards, problem- and idea-selection rubrics, and team portfolios.
Prototyping materials	A classroom Curiosity Box containing low-cost construction materials; Year 2 added batteries, circuit boards, sensors, LEDs, motors, propellers, and buzzers.
Feedback and iteration	Co-team feedback, whole-class feedback, user testing, individual reflection, team reflection, and cross-session improvement prompts.
Scheduling and culmination	Seven to nine weekly 90-minute sessions per block outside regular academic instruction, ending with team presentations of the problem, evidence, selected solution, prototype, and feedback.

Notes: Panel A summarizes the four curriculum blocks delivered during the two study years. Each block required teams to move from problem identification to investigation, idea generation and selection, prototyping, feedback, and presentation, with the degree of scaffolding declining over time. Panel B summarizes the classroom delivery protocol. Sources: Inqui-Lab student workbooks, student-leader handbook, team portfolios, session-detailing document, and Product Note.

Panel A: Guided



Panel B: Semi-open



Panel C: Independent

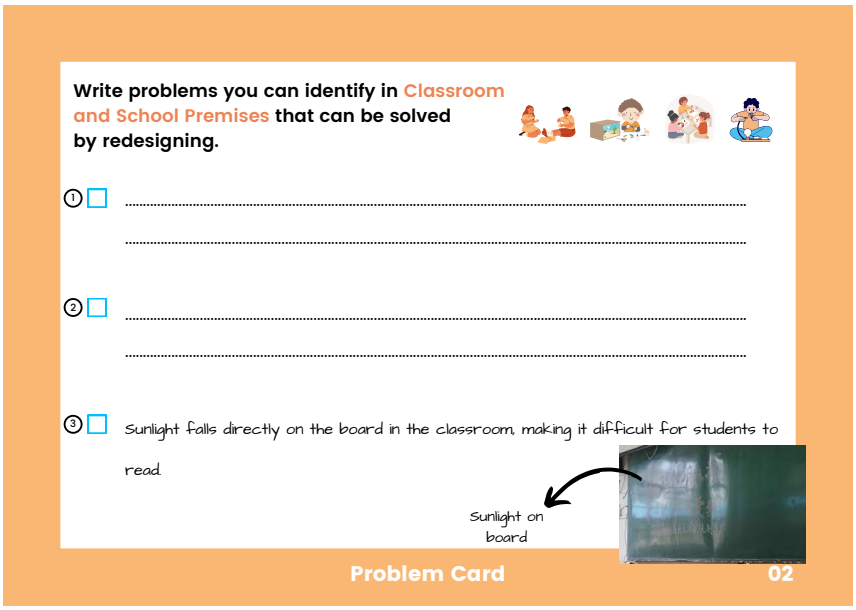


FIGURE B.1. PROGRESSIVE RELEASE OF AUTONOMY IN THINK & MAKE

Notes: Panel A shows a guided story-card comprehension task: the problem and possible interpretations are supplied. Panel B shows a semi-open design challenge: the object is assigned by a die roll and the user is specified, while the redesign is generated by the team. Panel C shows independent problem discovery: teams identify problems in their own school environment before ranking problems, generating and rating ideas, and planning a prototype. Across the three stages, students gain greater responsibility for defining the problem and solution; the progression involves more than a reduction in the amount of written guidance. Source: Inqui-Lab Foundation student workbooks.

C. The Innovation Challenge: Measurement and Materials

TABLE C.1—INNOVATION-S: ITEMS, CONSTRUCTION, AND VALIDATION

Item	Dimension	Adult-survey mean	Student mean	Inter-rater r
Panel A: items				
Offers a new solution to a problem	Novelty	4.34	3.64	0.26
Improves on an existing solution	Novelty	3.93	3.92	0.26
Combines existing solutions	Novelty	4.07	3.74	0.27
Transplants a solution across contexts	Novelty	3.71	3.68	0.24
Solves a problem affecting many people	Value-add	4.48	3.92	0.23
Improves the lives of intended users	Value-add	4.07	3.52	0.19
Can be accessed by those who need it	Value-add	4.55	4.13	0.23
Panel B: scale statistics				
Adult-survey respondents	44 answered at least one item; 40 complete on all seven			
Adult-survey Cronbach's α (complete cases)	0.485			
Student-sample scored submissions (two evaluators)	1,189			
Student-sample Cronbach's α	0.950			
First principal component, share of item variance	0.775			
Inter-rater correlation of the seven-item index	0.300 ($N = 1,181$ submissions scored by both evaluators)			
Treatment effect on the index, evaluator 1's scores only (SD)	0.225 (SE 0.066, $p = 0.001$)			
Treatment effect on the index, evaluator 2's scores only (SD)	0.179 (SE 0.065, $p = 0.007$)			

Notes: The items were developed from interviews with 14 adult innovators and a survey of 44 adults involved in developing products. The scale contains four novelty and three value-add items, each rated 1–5. Adult-survey means use respondents answering each item. Student means use submissions with complete mean item scores; inter-rater correlations use submissions with complete scores from both evaluators. Cronbach's α measures internal consistency (0.485 in the adult survey and 0.950 in the student sample); the index's inter-rater correlation is 0.300. For submissions with two ratings, the index averages the two evaluators' scores after standardizing within evaluator. The final two rows estimate treatment effects separately using each evaluator's scores among submitted teams, with the team-level specification in equation (3) and school-clustered standard errors.

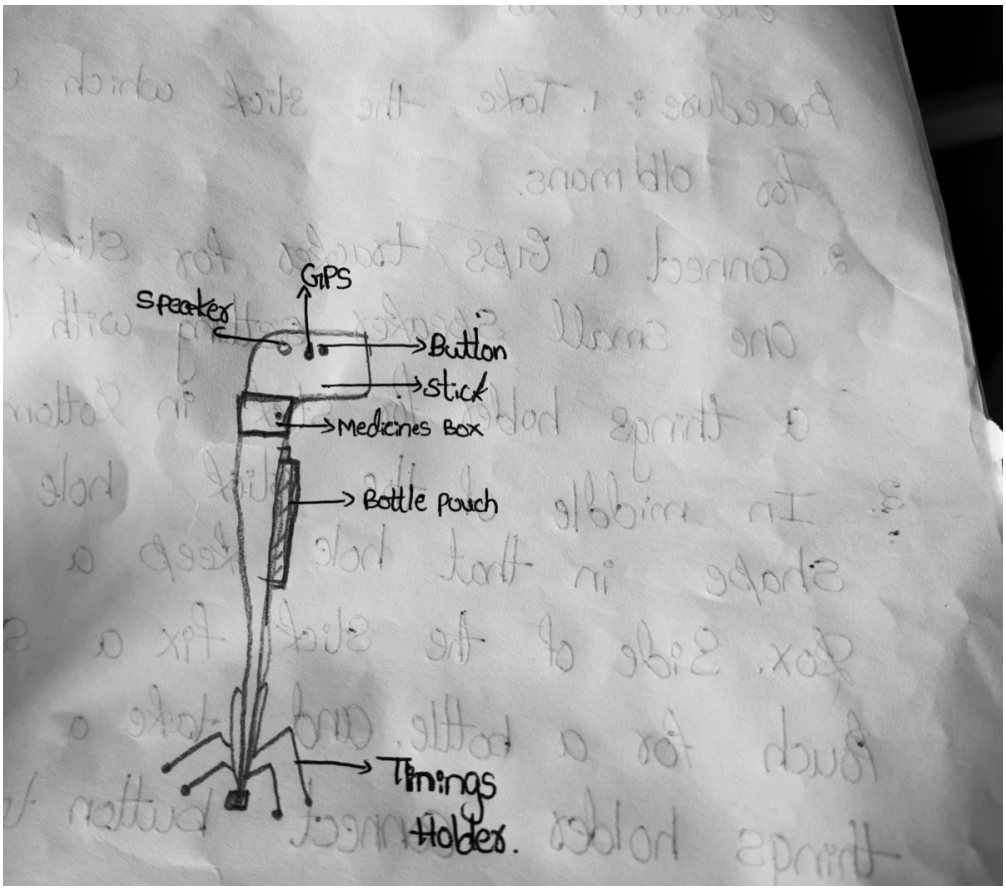
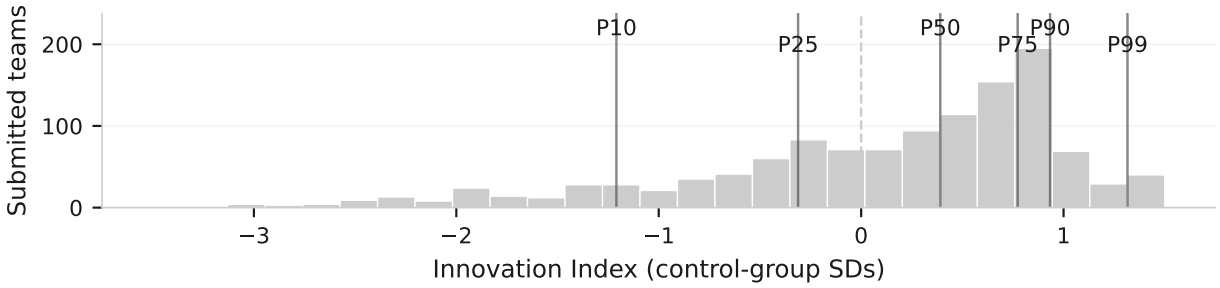


FIGURE C.1. A SUBMITTED IDEA DRAWING

Notes: The drawing accompanying a team's idea submission: a walking stick for elderly people who live alone and forget their medications, with a GPS tracker, a speaker for medication reminders, a medicines box, and a bottle pouch. The team was in a control school, and the submission is from the first (Year 1) cycle of the Innovation Challenge; the outcomes analyzed in this paper are Year 2 submissions, which used the same drawing component. In Year 2, submitting teams in both arms also provided an audio recording describing the problem, the solution, and its value; Year 1 pitches were transcribed in real time (Section IV.A); teams were not permitted to submit prototypes. The faint mirrored text is the team's written procedure on the reverse side of the page.



P10: index -1.21 SD

Problem: It is difficult for some companies to publicizr their brand successfully. They waste so much money on this.

Solution: To successfully publicize their brand, a helicopter is used. The brand is promoted via a helicopter with a speaker. As it goes around the town, a poster is also suspended so that people can look at it and...

P25: index -0.31 SD

Problem: how to be saved from monkeys, when we are classes, dorms, washrooms, we dont know if there is amonkey there or not, when we...

Solution: so we thought of a sensor which can be used anywhere for exampe dorm door or washroom door, if there is a monkey coming then the sensor buzzes and we will get intimitaded, because of the sound we will be...

P50: index 0.39 SD

Problem: Environment and natural resource: Water wastage due tonegligence towards motors. Often people forget to switch their water...

Solution: censor in water tanks and buzzers to indicate when the motor is full

P75: index 0.77 SD

Problem: Our problem is borewells. Many children are dying as they are falling into borewells while playing near them.

Solution: Our solution is to build it high and build steps in it along with placing sensors. The sensor will light up in blue and red colors in forests and villages respectively. This will alert the kids and blind...

P90: index 0.93 SD

Problem: when we use water, we release into drainage, but there are plastics, different types of toxics, other materials are also being...

Solution: so instead of directly releasing it , we can use this process where we recycle water and release them, we use 5 tanks, and there are different types of cleaning processes, in the first tank we use a process...

P99: index 1.32 SD

Problem: Blind, uneducated and old people find it difficult to access while travelling through a bus. They often get confused regarding...

Solution: When one goes to the bus station, the team proposes a sensor on each bus. This sensor gets activated once someone touches the stand on the bus and let people know where this it is headed. Additionally,...

FIGURE C.2. SUBMITTED IDEAS AT FIXED PERCENTILES OF RATED IDEA QUALITY

Notes: The histogram shows the distribution of evaluator-rated idea quality across submitted teams, measured by the combined novelty and value-add index in control-group standard deviations (Section IV.A). The six example submissions sit at the 10th, 25th, 50th, 75th, 90th, and 99th percentiles of that distribution; at each percentile the example is the submitted team with the most complete written record within a band around the percentile, selected without using treatment status, which is not shown. Each card reproduces, as submitted, the team's stated problem and proposed solution, together with the idea's index value; text beyond the card's space is truncated. The scale's construction and validation are reported in Appendix Table C.1.

Panel A. A finalist team presents its idea to a jury panel



Panel B. A finalist team demonstrates its prototype to jurors



FIGURE C.3. THE SHARK TANK FINALS

Notes: Photographs from the Shark Tank finals in Hyderabad. Panel A shows a finalist team presenting its idea to a jury panel; Panel B shows a finalist team demonstrating a working prototype to jurors. Faces are blurred to protect participants' privacy.

D. Skill Tasks: Design, Scoring, and Process Evidence

A. The Incentivized Search Task

TABLE D.1—DESIGN AND PAYOFF STRUCTURE OF THE INCENTIVIZED SEARCH TASK

	Fresh juice	Pani puri	Ice cream
Panel A: stand parameters			
Maximum hourly profit (V_s, Rs)	120	180	250
Penalty per unit of distance (λ_s, Rs)	10	20	30
Optimal location	Market	School	Homes
Optimal working hours	13	12	14
Optimal menu items	6	8	7
Optimal selling price (Rs)	15	13	14
Share of configurations with zero profit	0.257	0.540	0.633
Panel B: game constants			
Rounds	12		
Choice space	3 stands × 3 locations × 10 hours × 10 menu sizes × 10 prices = 9,000		
Feedback per round	own profit; choices and profits from completed rounds		
Demonstrated example (round 1)	fresh juice at the market, 5 items, 9 hours, Rs 19: Rs 30 per hour		
Display from round 2	student's best configuration and profit so far		
Input selection	round 1: active selection; rounds 2–12: previous choices pre-filled		
Prize	a geometry box for the highest single-round hourly profit in the class		

Notes: Panel A reports each stand's payoff parameters in the units students faced. Profit in a round equals the stand's maximum hourly profit minus the penalty rate times the sum of absolute deviations between the chosen inputs and the optimal inputs across the four dimensions (location, working hours, menu items, selling price), floored at zero (equation 1); for the penalty, working hours, menu items, and price are each measured on ten-point grids and location on a three-point grid, so a one-step deviation costs the same on every dimension. The zero-profit share is the fraction of the stand's 3,000 configurations returning zero profit and is computed from the payoff function. None of these parameters was disclosed to students. Panel B reports the game constants. The demonstrated example's profit of Rs 30 per hour is reproduced from the payoff function. Payoffs are deterministic: in the data, repeated plays of the same configuration always return the same profit.

TABLE D.2—SEARCH TASK: PROCESS MEASURES, ROBUSTNESS, AND OUTCOME CORRELATIONS

	Control mean	Effect	SE	p	N
Panel A: process measures and robustness					
Switched strategy after a zero-profit round (pp)	70.7	2.2	(1.2)	0.061	3,986
Switched strategy after a positive round (pp)	52.2	1.9	(1.9)	0.326	3,959
Difference: response to failure minus success (pp)	19.0	-0.5	(1.6)	0.773	3,557
Distinct configurations tried	7.965	0.247	(0.154)	0.113	4,388
Minimum distance, holding volume fixed	6.303	-0.181	(0.108)	0.098	4,388
Discovery, holding volume fixed	1.257	0.067	(0.035)	0.063	4,388
Distance relative to volume-matched random draws	-1.180	0.167	(0.106)	0.120	4,388
Panel B: correlations among the search outcomes					
	Initiation	Breadth	Discovery	Min. distance	Volume
Initiation	1.00				
Breadth	0.07	1.00			
Discovery	0.16	0.27	1.00		
Min. distance	-0.01	-0.20	-0.49	1.00	
Volume	0.07	0.78	0.45	-0.36	1.00

Notes: Panel A reports treatment effects on round-level process measures from the incentivized search task, estimated on the 4,388 students who completed it. A strategy switch is a change of stand or location relative to the previous round; switch rates are computed within student across rounds 2–12, conditional on the previous round's profit being zero (row 1) or positive (row 2), and row 3 is the within-student difference between the two responses. Distinct configurations tried counts unique combinations of all five choices across the 12 rounds. Rows 5 and 6 re-estimate the minimum-distance and discovery effects of Table 5 (Panel A) adding distinct configurations tried as a control. Row 7 is the difference between the expected minimum distance of uniform-random draws matched to the student's own number of distinct configurations (30,000 simulations per volume level) and the student's observed minimum distance, so positive values indicate proximity beyond matched random search; the row reports the treatment–control difference. Regressions control for randomization strata, school infrastructure, and student baseline covariates including the baseline exploration index, with indicators for missing values; standard errors, in parentheses, are clustered at the school level, and p-values use a t-distribution with degrees of freedom equal to the number of schools minus one. No significance stars are shown. Panel B reports pairwise correlations among the main-panel components and search volume.

B. The Skill Battery: Instruments and Scoring

TABLE D.3—SKILL BATTERY: ADMINISTRATION AND PSYCHOMETRICS

	Eyes test (RMET)	Matrix reasoning (ICAR)	Creativity tasks (PISA)	Sudoku task
Measured outcome	Emotional intelligence	Fluid intelligence	Idea generation; idea evaluation; written expression	Problem solving; challenge-seeking
Source design	Reading the Mind in the Eyes Test (Baron-Cohen et al. 2001)	International Cognitive Ability Resource (Condon and Revelle 2014)	PISA creative-thinking tasks (OECD 2024)	Incentivized easy-versus-hard choice (Alan, Boneva, and Ertac 2019), Sudoku in place of number-matching
Items or rounds	11 items	12 items	3 written tasks	5 rounds, four-by-four puzzles
Response format	4 labeled options, English and Telugu	4 options (reduced from 6 for tablet readability)	Open-ended text, English or Telugu	Easy-or-hard choice, then the puzzle; 2 minutes per round
Scoring	Two-parameter item response model	Two-parameter item response model	Task-specific rubrics applied by a large language model	Difficulty-weighted average score (hard $\times 1.5$); challenge-seeking = share of hard choices, rounds 2–5
Cronbach's α	0.46	0.79	Not applicable	0.84 (round scores)
Baseline form	Longer form, shortened to the most discriminating items	10-item form, recalibrated at endline	None; baseline exploration index is the control	None; baseline conscientiousness and fluid intelligence are the controls
Adaptation and piloting	Stimuli replaced for the Indian context; piloted with 150 out-of-sample students	Item order randomized within the module	Responses grammar-corrected and translated before rubric scoring	Hard puzzles have 50 percent more empty cells; round 1 imposed hard for a random half

Notes: The table documents administration and scoring for the four skill-battery instruments of Section IV.C; the incentivized search task is documented separately in Appendix Tables D.1 and D.2. All instruments were administered on individual tablets within the endline battery, with module order randomized across students; both item-based tests also randomized item order. Cronbach's α is computed from the analysis sample's item responses (binary items, so α is the Kuder-Richardson coefficient; for the Sudoku task it is computed over the five round scores); the creativity tasks are three distinct tasks rather than a scale, so no α applies. Item-response thetas and all scores are standardized to the control group. The large-language-model scoring of the creativity tasks is detailed in Appendix Table D.4. Skill-to-instrument mapping and sources appear in Table 3.



FIGURE D.1. SAMPLE ITEM: READING THE MIND IN THE EYES TEST

Notes: Each item presents a photograph of the eye region of a face; the student selects, from four labeled options presented in English and Telugu, the emotion that best describes what the person is feeling or thinking. The endline form has 11 items, administered on individual tablets with randomized item order and scored with a two-parameter item response model. The original stimuli (Baron-Cohen et al. 2001) were replaced with photographs appropriate to the Indian context, and the adapted instrument was piloted with 150 out-of-sample students before deployment. Administration and psychometrics for the full battery appear in Table D.3.

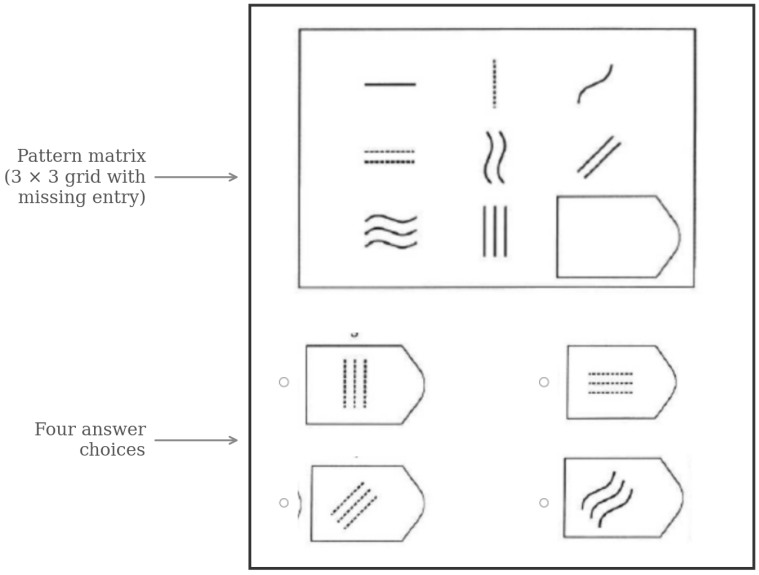


FIGURE D.2. SAMPLE ITEM: ICAR MATRIX REASONING

Notes: Each item presents a three-by-three grid of geometric patterns with one cell missing; the student completes the pattern from four options, reduced from the source battery's six for readability on tablet screens. The endline form has 12 items of increasing difficulty from the International Cognitive Ability Resource (Condon and Revelle 2014), administered with randomized item order and scored with a two-parameter item response model; a 10-item form was administered at baseline. Administration and psychometrics for the full battery appear in Table D.3.

TABLE D.4—PISA-ADAPTED CREATIVITY TASKS: SCORING RUBRICS AND EXAMPLE RESPONSES

Score	Rubric	Example Student Response	%
Panel A: Written Expression (Creative Story) (scored −1 to 5)			
-1	Non-interpretable or single word	*"Thdkfmgmgjyyigjgkg"*	21.8
0	Partial phrase or more than one word	*"The first image teaches the second image, then both teach together."*	8.2
1	One sentence with partial meaning	*"This is a picture with two images. There are arrows for some reasons."*	5.6
2	Multiple legible sentences without coherent idea	*"Tomorrow, I am going outside for shopping, and I want a toss. My friends have a toss, and my mom is very good."*	4.9
3	One meaningful sentence with image keywords (dice, globe, world, arrow, etc.)	*"Once upon a time on Earth, they were playing Ludo. One member won, and the other three members lost."*	25.4
4	Multiple sentences with partial/incomplete idea; no image keywords	*"Once upon a time, there was a girl named Niha and her friend. She was going to the office, and she would have a great..."*	14.7
5	Multiple sentences incorporating image keywords	*"In one universe, there was a planet named Earth. A big spaceship came and crashed into Earth, causing it to explode."*	19.5
Panel B: Idea Evaluation (Torch vs. Wheelchair) (scored 0 to 3)			
0	Non-interpretable or single word	*"Handicraft people are using bicycles."*	58.8
1	States preference or understanding without giving a reason	*"The wheelchair is very difficult for someone with only one hand and one leg; it is not suitable for such people."*	31.0
2	Gives a reason but does not address cost, impact, ease, or user needs	*"If a disabled person needs to go somewhere, they can fold the chair and take it along with them."*	5.6
3	Gives a reason about cost, help to users, impact, ease, or creative merit	*"With crutches, we can move very easily, but a torch can only be used at night."*	4.5
Panel C: Idea Generation (Bicycle Improvement) (scored 0 to 4)			
0	Non-interpretable or unrelated to bicycles	*"Gfkv"*	64.9
1	Mentions bicycle parts without suggesting improvement	*"I will improve the pedal."*	10.6
2	Very common features (lights, bell, brakes, basket, reflectors)	*"I will add a light to the bicycle."*	7.4
3	Common improvements beyond standard (electric motor, gears, suspension, speedometer)	*"I will add a comfortable seat."*	10.3
4	Unique or innovative (solar panels, distance sensor, auto-balancing, puncture-proof tires, energy-harvesting pedals, crash-alert notification, built-in first-aid kit)	*"I will add a distance sensor to the bicycle for crossing the road."*	6.9

Notes: The table reports task-specific rubrics and example responses for three PISA-adapted creativity tasks (OECD 2024). Percentages use 4,388 scored students in Panels A and B and 13,164 scored ideas in Panel C (three per student); non-interpretable responses enter at the bottom score level. Panel A scores a story about an image of dice, a globe, and an arrow from −1 to 5. Panel B scores the written justification of an allocation of Rs 1,000 between two inventions from 0 to 3. Panel C scores each proposed bicycle improvement from 0 to 4; the student's outcome sums the three scores. Responses were grammar-corrected and translated to English using GPT-4o (temperature 0.3), then scored against the task-specific rubric (temperature 0). Examples are students' responses after grammar correction and illustrate the score levels; Section IV.C describes the scoring procedure. Administration details appear in Table D.3.

E. Academic Outcomes

TABLE E.1—COMPOSITION OF REPORTED WEEKLY TIME

Outcome	Treatment effect	SE	Sharpened q	N
Time on school periods (% of reported time)	-0.18	(0.87)	[0.504]	4,388
Time on homework (% of reported time)	-0.36	(0.45)	[0.269]	4,388
Time on sports (% of reported time)	0.79**	(0.38)	[0.068]	4,388
Time on the Innovation Challenge (% of reported time)	1.31***	(0.38)	[0.005]	4,388
Time on other activities (% of reported time)	-1.53**	(0.76)	[0.068]	4,388

Notes: The table reports treatment effects on the five activity shares of students' reported weekly time, in percentage points, for the 4,388 students who completed the endline assessment. Shares come from a five-activity weekly hours grid (school periods, homework and study, sports, the Innovation Challenge, and other activities), with each activity expressed as a share of the student's own total reported hours; grids totaling 80 or more hours per week (28.3 percent of students) had all five shares replaced by the rounded mean shares of unflagged classmates in the same school section, and this replacement flag is balanced across arms ($p = 0.73$). Because the five shares sum to 100 for every student, the five treatment coefficients sum to approximately zero: the table reports treatment effects on the allocation of reported weekly time across activities. Students could describe the other-activities category in a free-text field; entries in both arms include dance, singing, drawing, and reading. Every regression controls for randomization strata, school infrastructure, and individual covariates (age, socioeconomic index, computer skills), together with baseline mathematics and science scores and their missing-value indicators. Standard errors, in parentheses, are clustered at the school level, the level of treatment assignment; significance stars refer to the clustered p-value. Values in square brackets are Anderson sharpened false-discovery-rate q-values, computed within the five-share family. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

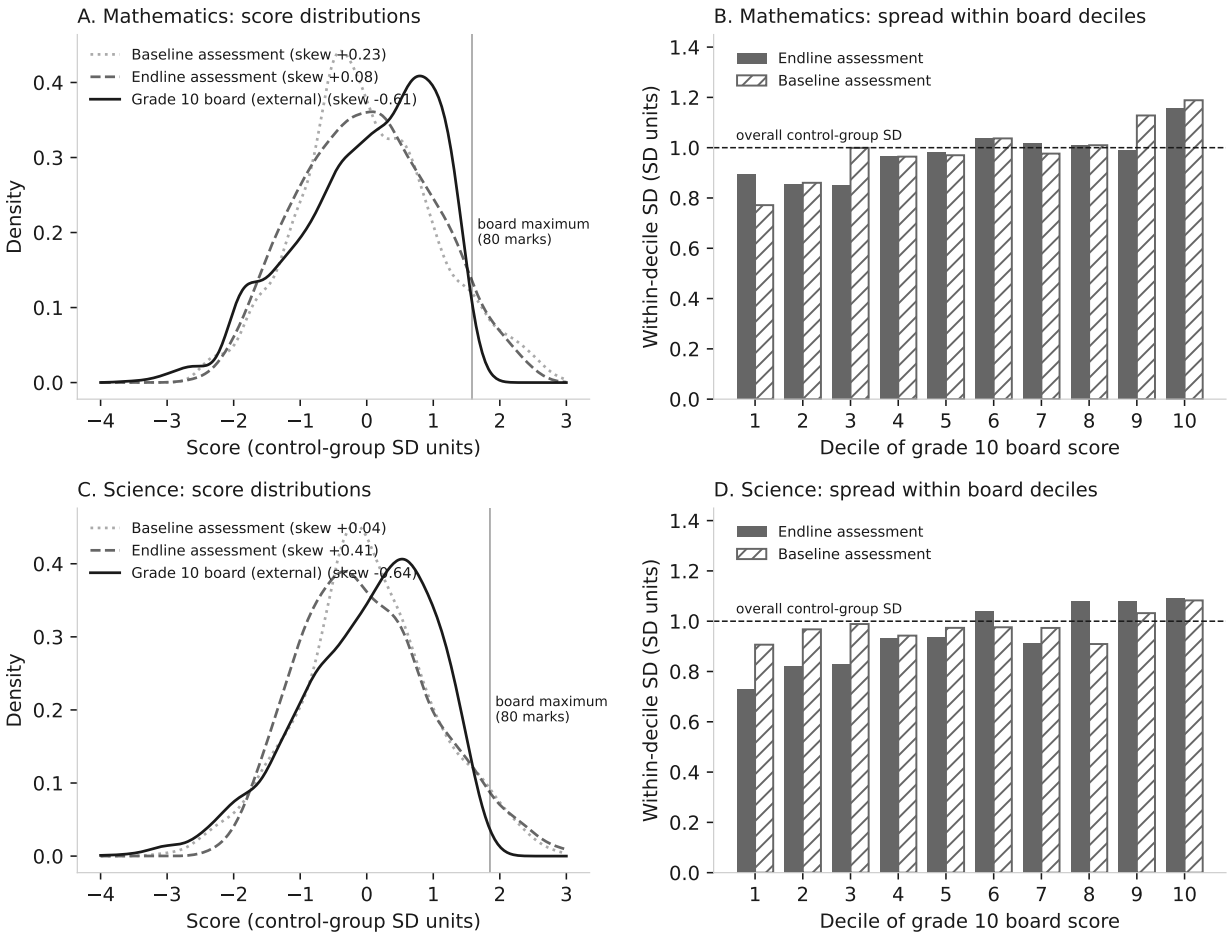


FIGURE E.1. CEILING COMPRESSION IN THE GRADE 10 BOARD EXAMINATIONS

Notes: The figure compares, among control-group students, the two study-administered assessments in each subject (baseline and endline, different item sets administered three academic years apart, each scored by a two-parameter item-response model) with the externally graded component of the grade 10 board examination in the same subject; each score is standardized to the control group ($N = 2{,}123$ students with all three scores in each subject). Panels A and C show score distributions; vertical markers identify the board examination maximum of 80 marks. Panels B and D plot the within-decile standard deviations of baseline and endline assessment scores by grade 10 board-score decile. The horizontal reference line marks the full control-group standard deviation.

F. Student Feedback at Midline

At the close of the Year 1 (midline) tablet assessment, students in both arms answered one open-ended prompt, administered in English and Telugu: "Please write your thoughts about Inqui-Lab's Think and Make Program, through which you learned to work on two new ideas and products this year. What have you learned from this program?" Table F.1 summarizes the responses.

TABLE F.1—OPEN-ENDED STUDENT FEEDBACK AT MIDLINE: RESPONSE FUNNEL, VALENCE, AND THEMES

	N	Share
Panel A. Response funnel		
Students assessed at midline	4,620	
Wrote a non-empty comment	4,612	
Decipherable comment	4,432	
Substantive (at least one theme)	4,315	
Panel B. Overall valence (share of decipherable comments)		
Positive	4,004	90.3%
Neutral	412	9.3%
Negative	16	0.4%
Panel C. Content themes (share of substantive comments; multi-label)		
Enjoyment / affect	3,480	80.6%
Program-name mention (primed by the question)	1,079	25.0%
Assessment session (game, quiz, tablet, marks)	871	20.2%
Generic learning claim	801	18.6%
Gratitude	587	13.6%
Problem-solving process language	564	13.1%
Other themes	727	16.8%

Notes: The sample pools both study arms (the prompt named the program to both, and control students had themselves taken part in the Ideas Challenge); the table is descriptive and not a treatment-effect comparison. Students left these comments in the final block of the Year 1 (midline) tablet assessment, in response to a single open-ended prompt (Appendix F). Panel A reports the response funnel: a comment is "decipherable" if it carries intelligible content (single letters and keyboard fragments are excluded) and "substantive" if it expresses at least one of the ten coded themes. Panel B shares are of decipherable comments. Panel C shares are of substantive comments; a comment may express more than one theme, so the shares sum to more than 100 percent, and *Other themes* collects the remaining, lower-frequency codes (a comment counts there if it carries at least one). Each comment was coded by a large language model, blind to treatment assignment, against a fixed codebook; the author read a random sample of the coded responses against the codebook. The codebook, the prompt, and the coded corpus are in the replication package.

The responses below, from students who received the curriculum, are lightly edited for readability and, where originally in Telugu, translated; the verbatim originals are in the replication package.

- *Problem-solving.* "I liked the Think and Make program and the Inqui-Lab Foundation because they are very useful for the development of the country. Children can take more initiative in solving problems. I am a student leader, and I gained a lot of experience in teaching problem-solving methods."

- *Making and community.* "I have learnt a lot from this Inqui-Lab. In this Think and Make program we made two prototypes. I learnt new things, and how to solve the problems around me and how to overcome them."

- *Change in school (translated from Telugu).* "I liked it very much because we worked as a group and had team discussions. Changes came to our school too: we started using steel bottles, and because of Inqui-Lab we stopped wasting resources. Thank you for this opportunity."

- *Transfer to daily life.* "I like this program because I think it is very innovative. I solved so many problems in the Think and Make program. This is really useful in our life: I learnt that if we have any problem in our surroundings, we can solve it."

- *Aspiration.* "I will find problems in our house, gurukulam, and school, and then find their solutions. Inqui-Lab's Think and Make program made me so happy. I want to become a scientist in my life. Thank you for this opportunity, sirs and madams."

G. Deviations from the Pre-Analysis Plan

The experiment was registered at the AEA RCT Registry as AEARCTR-0009696 in August 2022. The assessment instruments were uploaded in January 2024, before endline data collection; the registry's structured outcome fields were not updated. Changes from the registration are listed below.

- **Measurement.** The registered Innovators DNA survey was not administered.
- **Measurement.** The risk-taking task (Gneezy and Potters 1997) and the Big Five inventory were administered at baseline and midline but dropped from the endline battery: the risk task produced almost no variation at midline, the inventory is self-reported, and the assessment time was reallocated to performance-based tasks.
- **Measurement.** The three PISA-adapted creativity tasks and the timed puzzle task were not named in the original registration but were included in the uploaded instruments.
- **Measurement.** The puzzle task was intended to measure perseverance; because the task provides no per-round feedback, that construct is not identified, and the paper reports the difficulty-choice margin (challenge seeking) instead.
- **Measurement.** The registered exploration instrument (the "Lemonade Test" of Ederer and Manso) was adapted for tablet administration with three food stands, revised choice variables, and a class prize based on the highest profit achieved in any round. The registered construct, experimentation and tinkering, is preserved and is the outcome family of Table 5, Panel A.
- **Outcomes.** The registered team outcome "community feedback" was not collected at endline.
- **Outcomes.** Grade 10 board examination results were added after registration as administrative data under a separate approval by the Stanford University IRB (protocol 79854, approved April 4, 2025, with waivers of consent and assent); they are reported as the academic boundary outcome (Table 6).
- **Analysis.** Heterogeneity in team outcomes is pre-specified in the trial registry; the baseline-score splits are exploratory. Social-network measures and the registered team-composition analysis are reserved for a companion paper; teams were randomized within schools in both arms, which constitutes an independent experiment.