Can Innovation be Taught in Schools? Experimental Evidence from India*

Saloni Gupta

Latest version: github.com/salonigupta435/jmp

Abstract

As automation substitutes for routine cognitive tasks, returns to creativity, analytical reasoning, and problem-solving are rising. But can these skills be taught at scale through public education? I evaluate a two-year innovation curriculum using a cluster randomized trial across 80 government schools and 5,600 students from marginalized communities in India. I develop a measurement toolkit combining an incentivized exploration task, large-scale creativity and problem-solving assessments, and an externally evaluated team idea competition. The curriculum increases a creative problem-solving composite by 0.12 SD, with the largest effects on idea generation (0.31 SD), problem solving (0.14 SD), and fluid intelligence (0.11 SD). Effects are positive across the ability distribution but exhibit a gradient: for the highest-ability quintile, training raises both creative problem-solving (0.26 SD) and mathematics (0.22 SD), while below-median students develop above-median creative problem-solving at 7 percentage points higher rates. Innovation content varies by identity: girls' teams disproportionately address social welfare problems; rural students target agriculture. At $40 per student annually, the program develops these skills without reducing academic achievement, even in the most resource-constrained government schools.

JEL Codes: I21, I25, O31, O15, J24
Keywords: Innovation, creativity, education, RCT, India, skills

I. Introduction

School systems worldwide sort students on what their examinations can capture: content recall, procedural fluency, formulaic application (World Bank, 2018; Muralidharan, 2024). But returns to the routine skills these tests reward are falling, while returns to creativity, analytical reasoning, and problem-solving are rising (Hermo et al., 2022; Deming and Silliman, 2024; Deming, 2017). When schools fail to develop these skills, acquisition defaults to exposure outside the classroom. Evidence from the United States illustrates the consequence: children from low-income families are ten times less likely to become inventors than those from

*Gupta: Brown University, Department of Economics (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, 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 (Protocol #22-352, approved 07/07/2022).

the top percentile, even at identical test scores, a gap Bell et al. (2019) attribute to differences in exposure rather than ability.

Artificial intelligence is widening the gap between what education systems produce and what labor markets demand: routine cognitive skills are now those most rapidly being automated (Autor et al., 2003; Acemoglu et al., 2026). Entry-level employment in AI-exposed occupations has already declined 13 percent since 2022 (Brynjolfsson et al., 2023). PISA mathematics scores, which test applied reasoning in real-world contexts rather than routine procedures, fell 15 points across OECD countries between 2018 and 2022, a worrying trend at a time when labor market demand for non-routine skills is accelerating (OECD, 2023). Education systems built to reward routine cognition have not yet caught up to these new demands.

The consequences are visible in labor markets worldwide. Sixty-three percent of employers across 55 economies identify skills gaps as their primary barrier to business transformation, ranking analytical and creative thinking among the capabilities most in demand and hardest to find (World Economic Forum, 2025). In India, fewer than half of graduates are deemed employable despite holding college degrees (Wheebox et al., 2023; Aspiring Minds, 2019; Mercer Mettl, 2025). The skills in short supply are precisely those whose returns are rising, whose value increases as they combine with AI, and that could broaden who gains access to innovation pathways.

This paper tests whether a structured curriculum can develop creativity, analytical reasoning, and problem-solving within government schools serving low-income communities. The OECD Oslo Manual defines innovation as a new or improved product or process made available to potential users (OECD and Eurostat, 2018). The curriculum adapts this process for a school setting in which students are otherwise exposed to lecture-based instruction, procedural mathematical problem-solving, and rote science curricula that rarely connect classroom content to real-world problems or contemporary technologies. For example, a team identifies contaminated drinking water in their school and builds a clip-on indicator for water bottles that signals whether the water is safe, a product with clear users, a real need, and no existing solution in their setting. I evaluate this curriculum in a cluster randomized trial across eighty government residential schools in Telangana, India, following 5,642 students over two academic years.

There is growing evidence that higher-order skills are malleable through structured intervention. Experimental work in schools has shifted individual dispositions such as patience, grit, curiosity, and growth mindset (Alan and Ertac, 2018; Alan et al., 2019; Alan and Mumcu, 2024; Yeager et al., 2019; Bettinger et al., 2018). Inquiry and problem-based pedagogy has improved learning outcomes across field experiments in Latin America and Uganda (Bando et al., 2019; Ashraf et al., 2025a). Structured programs for adolescents in Sub-Saharan Africa have developed business and negotiation skills with effects persisting a decade later (Chioda et al., 2026; Ashraf et al., 2025b).

Translating these findings into programs that work sustainably within government school systems is an open challenge: teachers are overburdened and evaluated on syllabus completion, scheduling is rigid, and additional instructional demands face institutional resistance. The curriculum evaluated in this study was designed around these constraints. It is student-led: teams work from a structured manual and materials kit, with trained student leaders facilitating sessions and teachers serving as enablers rather than delivering content. It runs during after-school hours, displacing one weekly homework period rather than classroom instruction. Residential schools offer scheduling flexibility, but the program requires only one weekly session, a demand that non-residential schools could accommodate by displacing a single activity or revision period.

The curriculum was designed by Inqui-Lab Foundation, an education non-profit founded by engineers who became educators, that refined it over five years of implementation in these schools before this study began. The non-profit provides beginning-of-year training to student leaders and periodic monitoring with regular visits to troubleshoot and observe fidelity, while schools run the program independently, a delivery model consistent with evidence that light-touch external quality assurance sustains program effects at scale (Muralidharan and Singh, 2025). The curriculum builds higher-order skills through a team process: teams identify real problems from their own environments, investigate causes, generate and evaluate solutions, build

prototypes, and present their work to peers and teachers. This cycle is repeated over two academic years following students through Grades 8 and 9 across approximately forty-five contact hours. Generating and evaluating ideas builds creativity. Investigation and root cause analysis build analytical reasoning. Iterative prototyping builds problem-solving.

The measures are organized around three questions. First, do component skills improve? An incentivized exploration task adapted from Ederer and Manso (2013) captures iterative experimentation: how students search for solutions through structured tinkering. Matrix reasoning items from the International Cognitive Ability Resource (ICAR) measure fluid intelligence (Condon and Revelle, 2014). Sudoku puzzles of varying difficulty measure structured problem-solving and persistence within a well-defined problem space (Newell and Simon, 1972). Three open-ended tasks adapted from the PISA Creative Thinking framework measure written expression, idea evaluation, and idea generation (OECD, 2024). Second, does applied innovation improve? In a year-end ideas competition, both treatment and control teams developed solutions to real community problems. Each team recorded a structured pitch in their own voice, preserving the idea as students conceived it rather than filtering it through enumerator interpretation. These audio recordings were transcribed and scored independently by two external evaluators, blind to treatment status, on a validated innovation index. Third, does innovation training displace academic learning? Endline mathematics and science assessments, calibrated using two-parameter IRT models with high test information across the ability distribution, provide a precise test. Administrative records from the Telangana Class X board examinations, taken one year after the intervention ended, provide an independent check, though ceiling compression in these exams limits power to detect effects among students in the top quintile of baseline ability.

The assessment design makes it possible to measure higher-order skills within the infrastructure that government schools already use. All measures are tablet-administered, with question and module order randomized across students so that neighbors encounter different items at any given time, reducing copying. Module-level timestamps support detection of rushed or disengaged responses. Because data are captured digitally, a large item bank could be pilot-tested and rapidly iterated before the main launch, improving the test information function across all modules. Because the entire battery runs on tablets, it can be deployed on devices that state governments across most developing countries have already placed in schools. School systems develop what their examinations can measure, and these instruments bring higher-order skills within reach of the same infrastructure.

The evaluation is set in one of the most resource-constrained public school environments in India: eighty single-gender government residential schools in Telangana serving exclusively students from Scheduled Caste and Scheduled Tribe communities, among India's most economically marginalized populations. I conduct a cluster randomized controlled trial across these schools (forty treatment, forty control), following 5,642 students over two academic years (2022–2024). If innovation-relevant skills can be developed in these schools, resource constraints are unlikely to prevent adoption elsewhere. Both treatment and control schools participated in an annual innovation competition; only treatment schools received the structured curriculum. The comparison isolates whether structured pedagogy develops these skills beyond what opportunity alone provides.

I report four main findings. First, a structured curriculum can develop the creative and analytical skills that education systems have struggled to produce within the constraints of public schooling. The innovation curriculum raises creative problem-solving by 0.120 standard deviations, a composite that is the equal-weighted mean of three pre-registered domains: creativity, analytical reasoning, and exploration. Idea generation increases by 0.313 SD, problem solving by 0.142 SD, and fluid intelligence by 0.111 SD. On the exploration domain, treated students discover higher-value outcomes (0.094 SD), search more broadly (0.077 SD), and converge closer to optimal solutions (0.086 SD), with the advantage compounding over rounds rather than fading, evidence that students internalize a process for approaching unfamiliar problems, not just the specific content of the training. The curriculum targets each of these skills through distinct stages of a repeated innovation cycle (Figure 1).

Second, the curriculum develops creative problem-solving across the entire ability distribution, and for the top quintile of baseline ability, it raises mathematics performance as well. The ability gradient is clear: effects on idea generation are more than twice as large for above-median students (0.457 SD versus 0.198 SD, $p_{\text{diff}} = 0.012$), and the academic complementarity is confined to the top quintile, where the curriculum raises both creative problem-solving (+0.258 SD) and mathematics (+0.218 SD). Among academically below-median students, 7.3 percentage points more score above the median on creative problem-solving in treatment schools than in control schools—skills that, without structured exposure, would go undeveloped rather than merely undetected. The children in these schools come from families where parents work as daily-wage laborers in farming and construction. These results suggest that structured training, not just opportunity, is what develops creative problem-solving.

The program produces its largest effects in girls' schools, where idea generation increases by 0.504 SD compared with 0.095 SD in boys' schools ($p_{\text{diff}} = 0.010$). Because all eighty schools are single-gender, this comparison captures classroom composition alongside gender rather than gender in isolation. Effects are also larger in social welfare schools than tribal welfare schools (0.210 versus 0.058 SD, $p_{\text{diff}} = 0.023$). Social welfare schools enter with baseline academic ability 0.40 SD above tribal welfare schools, and given the ability gradient documented above, this differential likely reflects the concentration of higher-ability students rather than a pure institutional effect.

Third, component-skill gains translate into applied innovation. In a year-end ideas competition scored blind by external evaluators ($N = 1{,}283$ teams), treatment teams score 0.223 SD higher on a validated innovation index, with the largest component effect on novelty (0.263 SD). Both treatment and control students had the same opportunity to compete; the difference is that treatment students had a structured process for identifying problems, generating solutions, and refining ideas. As external corroboration, top-performing teams competed in a live funding competition judged by industry professionals; eight of the ten teams receiving the largest allocations are from treatment schools. That industry judges, evaluating ideas without knowledge of treatment status, independently favored the same teams suggests the quality difference is legible beyond the research setting. Innovation content reflects students' lived contexts: girls' teams disproportionately address social welfare problems, tribal school teams target agriculture, and the curriculum shifts treated teams toward education-sector innovations. Sector choice varies more by demographics than by treatment status (χ^2 $p = 0.001$ for gender, $p < 0.001$ for location and school type), consistent with evidence that inventor identity shapes the direction of innovation (Koning et al., 2021; Einiö et al., 2023).

Fourth, a natural concern for policymakers is whether innovation training comes at the cost of academic learning. The endline academic composite shows a precisely estimated null effect (0.011 SD, SE = 0.056). A quintile decomposition confirms that academic effects are null for the bottom four quintiles of baseline ability. Administrative data from the Telangana State Board Class X examinations, externally administered approximately one year after the intervention ended, confirm no academic harm: treatment effects on mathematics (0.112 SD), science (−0.006 SD), and total score (0.135 SD) are all statistically indistinguishable from zero. Language subjects show marginally positive effects (0.149 to 0.199 SD), consistent with the curriculum's emphasis on articulation. The overall pattern is clear: the program develops creative problem-solving without sacrificing academic performance at any point in the ability distribution, and for the top quintile, it improves both.

This paper contributes to three strands of literature. First, a growing body of work documents that who becomes an innovator depends on exposure rather than ability. Bell et al. (2019) show that children from low-income families are far less likely to patent even at identical test scores, a gap they attribute to differences in childhood exposure to innovation. Aghion et al. (2017) find that parental income strongly predicts who becomes an inventor in Finland, even conditional on ability. Toivanen and Väänänen (2016), exploiting a Finnish polytechnic reform, demonstrate that access to technical education causally increases patenting. Even within higher education, Biasi and Ma (2022) show that access to frontier knowledge is unequal: universities enrolling socioeconomically advantaged students teach content closer to the research frontier, and their

graduates produce more patents and earn more. These papers establish that the ability to innovate is broadly distributed but access to environments that develop it is not. Moreover, Koning et al. (2021) and Einiö et al. (2023) show that inventor identity shapes the direction of innovation: who gets to innovate determines what problems get solved. This paper provides experimental evidence that structured training within government schools can develop the creative and analytical skills that this literature identifies as undersupplied among disadvantaged populations, with gains emerging across the entire ability distribution and, for the top quintile of baseline ability, raising both creative problem-solving and mathematics performance. A directionally consistent pattern appears at the educational stage, with sector choice in the ideas competition varying systematically by gender and community context.

Second, Deming (2022) identifies an asymmetry at the center of human capital research: the technology for producing foundational skills like numeracy and literacy is well understood, but the technology for producing higher-order skills such as problem-solving and teamwork is not. A broad literature establishes that non-cognitive and social skills predict labor market outcomes, educational attainment, and life success as strongly as cognitive ability (Heckman and Kautz, 2012; Lindqvist and Vestman, 2011; Deming, 2017; Edin et al., 2022). Within schools, Jackson et al. (2020) show that effects on socioemotional development are as large as effects on test scores and predict long-run outcomes including reductions in arrests. Yet measurement of non-test-score school outputs has relied on self-report surveys susceptible to reference bias (West et al., 2016). Behavioral measures of cognitive skills at scale remain scarce. This paper addresses both sides of the problem. It validates a multi-method measurement strategy for creative problem-solving, spanning incentivized behavioral tasks, LLM-scored open-ended assessments, and blind-evaluated team competitions, all tablet-administered and scalable to government school settings. Treatment effects appear consistently across measures that differ in structure, incentives, and scoring, suggesting the gains are unlikely to reflect any single measurement artifact.

Third, a growing experimental literature establishes that targeted training can develop economically relevant skills, from individual dispositions in schools such as patience, grit, curiosity, and growth mindset (Alan and Ertac, 2018; Alan et al., 2019; Alan and Mumcu, 2024; Yeager et al., 2019; Bettinger et al., 2018) to business and negotiation skills for adolescents and adults with effects persisting a decade (Chioda et al., 2026; Ashraf et al., 2025b). This paper develops multiple cognitive skills simultaneously through a single curriculum integrated into government schools, with effects that translate into applied innovation—an outcome more directly linked to economic growth and productivity than the individual dispositions studied in prior work.

The policy stakes of developing innovation skills in schools are immediate: governments across income levels are actively scaling innovation education programs, committing substantial public resources without evidence on which approaches work or what they produce. India's Atal Tinkering Labs and School Innovation Marathon, South Korea's STEAM Pioneer Schools, Singapore's Applied Learning Programme, and the United Kingdom's Creative Partnerships represent major public investments in this space (NITI Aayog, 2020; Ministry of Education, Atal Innovation Mission, NITI Aayog, 2024; Delhi Government, 2019; Kang, 2019; Ministry of Education, Singapore, 2013; Creativity, Culture and Education, 2012). Several countries have gone further, embedding creativity and innovation into national curricula, including India's National Education Policy 2020 (Finnish National Board of Education, 2016; Australian Curriculum, Assessment and Reporting Authority, 2013; Ministry of ICT and Innovation, Rwanda, 2019; Ministry of Education, India, 2020). The scale of these commitments makes the absence of rigorous evidence on their effectiveness striking. The program evaluated here costs approximately forty dollars per student per year and produces measurable gains in creative problem-solving without reducing academic achievement.

The remainder of the paper is organized as follows. Section II describes the intervention and setting. Section III presents the experimental design and data. Section IV details the measurement strategy. Section V reports the main results. Section VI examines heterogeneous treatment effects. Section VII discusses mechanisms and external validity. Section VIII concludes.

II. Context and Intervention

The setting is Telangana's government welfare residential school system. The state operates 248 such schools under two management types—Social Welfare and Tribal Welfare—serving exclusively students from Scheduled Caste and Scheduled Tribe communities, among the most economically marginalized in India. Students board at the schools; most come from rural households where parents work as daily-wage laborers in farming or construction. Since 2017, the Inqui-Lab Foundation has delivered a two-year structured problem-solving curriculum—Think & Make—in these schools, in partnership with the Government of Telangana.[1]

The program comprises four instructional blocks spread across two years, with each block containing 7–8 sessions of approximately 90 minutes—roughly 45 contact hours in total. Students work in permanent teams of three to five (typically four), progressing through a repeated five-stage cycle. Sessions occur weekly during after-school hours—these are residential schools—and displace homework time. They are facilitated by trained student leaders; teachers provide support by advising on student ideas and helping organize end-of-unit presentations in front of the school assembly.[2]

Each block follows the same five-stage process—Find, Learn, Ideate, Make, Present (FLIMP). Each cycle asks students to: discover a real-world problem by observing and interviewing stakeholders (perspective-taking, emotional intelligence); investigate causes using analytical tools such as cause–effect diagrams (systematic reasoning, fluid intelligence); generate and evaluate multiple solutions under constraints (divergent thinking, creativity); build and test a physical prototype from low-cost materials (problem solving, grit); and present their solution to peers and respond to feedback (communication, metacognition). The repeated practice of these cognitive demands over four cycles, with progressively fading support, is how the program is hypothesized to build both higher-order skills and a structured approach to problem solving.[3]

The program employs a progressive release of autonomy across the four blocks. In Year 1, both blocks use guided story cards featuring model problem-solvers, comprehension checks, and structured evaluation rubrics. Year 2 Block 1 transitions within a single block: guided story cards give way to semi-open quick challenges (randomized design problems under time constraints), then to fully independent problem discovery where students identify, scope, and solve their own problems. Year 2 Block 2 escalates further by introducing electronic sensors and technology-integrated prototyping, while repeating the guided-to-independent progression. The program thus builds the FLIMP process as a transferable cognitive habit through deliberate practice with fading support.[4]

Figure 1 maps the curriculum to outcomes. The FLIMP cycle exercises specific higher-order skills at known dosages (Appendix Table A.2). Innovation ability—what students produce when they combine these skills with a structured problem-solving process—is the primary outcome, measured through a creative problem-solving (CPS) composite and applied innovation assessments. Academic performance serves as a test for spillover and displacement: the program contains no curricular content in examination subjects, so any positive academic effects represent cognitive transfer, while the program displaces study time, creating a negative mechanical channel.

Three features of the program sharpen the treatment-control contrast. First, the program contains no direct instruction in mathematics, science, or any examination subject. Students solve real-world design problems—chalk waste, paint drips, plastic cups, sensor-based devices—so any academic effects reflect cognitive transfer,

[1] India's School Innovation Marathon, administered by NITI Aayog's Atal Innovation Mission, reaches over 670,000 students across 54,000 schools but provides opportunity without structured training. AIM also operates over 10,000 Atal Tinkering Labs (ATLs) in schools nationwide, which provide physical infrastructure and some facilitated tinkering activities. The School Innovation Marathon, by contrast, is a voluntary competition open to all schools—including those without ATLs—and offers only minimal curriculum support; participation is not required.

[2] Appendix Table A.1 provides a complete session-level audit of all 30 sessions across the four blocks.

[3] Appendix Table A.2 maps each skill to its FLIMP stage(s), curriculum examples, and cognitive demands.

[4] Appendix Figure A.1 reproduces three workbook pages illustrating the guided, semi-open, and independent modes.

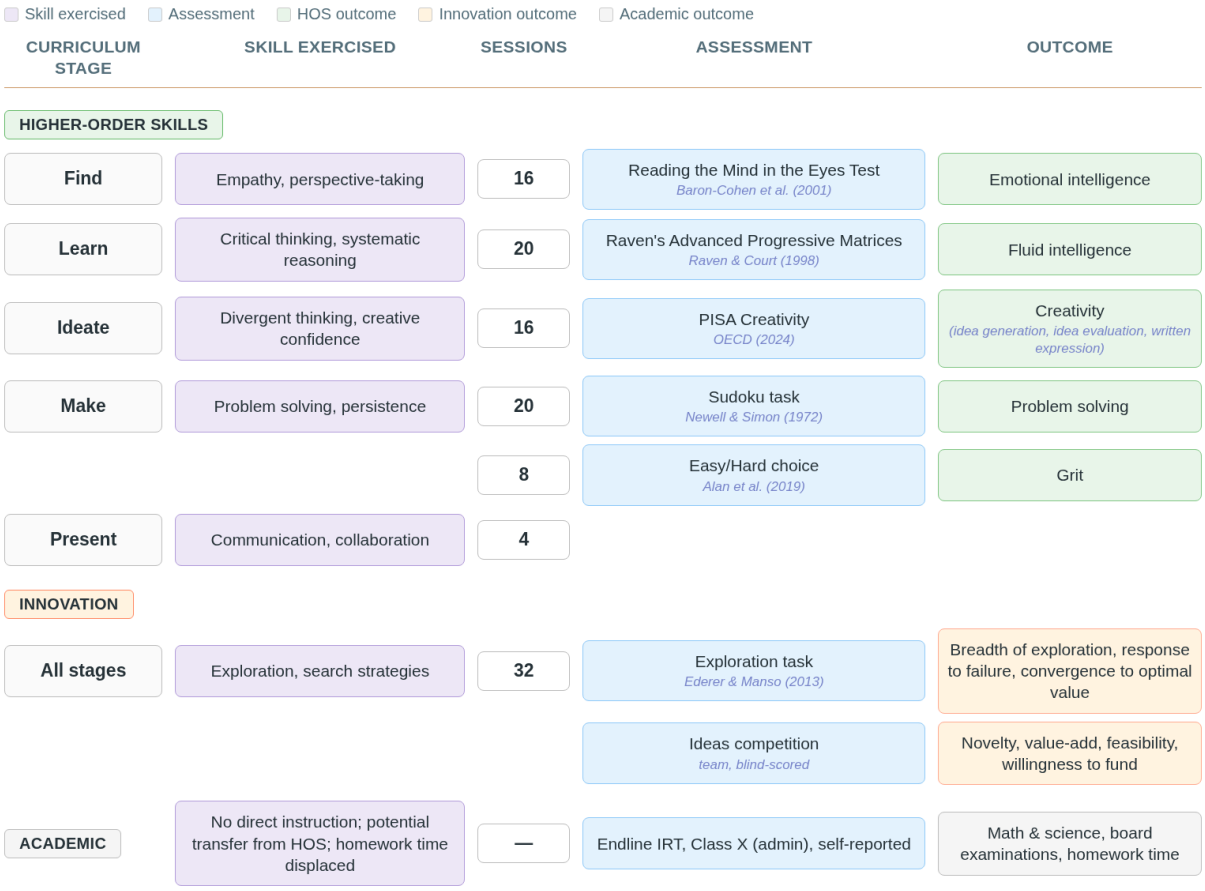


FIGURE 1. CURRICULUM AND OUTCOME MAPPING

Notes: The left panel maps the five FLIMP stages to the higher-order skills each stage exercises. The right panel connects those skills to the study's outcome domains. The intervention section documents the curriculum-to-skill links (left panel); the measurement section documents the skill-to-measurement links (right panel).

not content overlap. Second, the program is not teacher-delivered. Student-leader facilitators manage time and materials but do not lecture, grade, or correct, ruling out a "better teaching" channel. Third, sessions occur during after-school hours and displace homework time. Control schools participate in the same year-end Innovation Challenge—a competition in which teams pitch solutions to real problems, modeled on formats that have become culturally salient in India through *Shark Tank India*, which has fueled widespread student interest in entrepreneurship and innovation (Soni and Sharma, 2022)—but receive no structured curriculum. Control students were enthusiastic about the competition; the difference is that they lacked a systematic process for developing their ideas. Endline time-use data show that control students allocate similar shares of their non-school time to homework, sports, and innovation-related activities, confirming that the counterfactual is diffuse rather than concentrated in any single substitute activity.[5]

[5]Appendix Section D provides detail on the delivery model, facilitator training, and counterfactual conditions. Control-group time-use figures are from the individual endline survey ($N = 2{,}126$ control students).

III. Experimental Design and Data

The evaluation has two objectives: to test whether a structured innovation curriculum develops measurable cognitive skills, and to determine whether any such gains come at the cost of academic performance. I address both using a school-level cluster-randomized trial. From the 248 government welfare residential schools in Telangana, 80 were selected on the basis that no other externally run program was active at enrollment and that the implementing foundation could manage them operationally, ensuring that estimated treatment effects are not confounded by concurrent interventions. The selected sample spans all school types in the system—girls-only and boys-only, Social Welfare and Tribal Welfare, urban and rural—so that results speak to the full range of institutional settings rather than a narrow subset.

The 80 schools were stratified along three dimensions—school management (Social Welfare vs. Tribal Welfare), student composition (girls-only vs. boys-only), and location (rural vs. urban)—producing eight strata. Within each stratum, half of the schools were randomly assigned to treatment and half to control.[6] School-level assignment avoids within-school contamination: all Class 8 students in a treatment school receive the program, and all Class 8 students in a control school serve as controls. Because randomization occurred at the school level, no individual student data existed before the baseline survey. The analysis frame comprises all students enrolled in Class 8 across the 80 schools at the time of the baseline, totaling 5,642 students (2,835 treatment, 2,807 control) with approximately 70 students per school.

Figure 2 summarizes the experimental design. Data collection spans three years. All assessments were administered on individual tablets using the Qualtrics Offline App, which enabled survey delivery in schools without reliable internet connectivity. Each student received a tablet preloaded with the complete assessment battery in both English and Telugu. The order of assessment modules was randomized across students to mitigate survey fatigue and ordering effects, and to reduce opportunities for copying—students seated next to one another encountered different content at any given time. Each question appeared on a separate screen, further limiting cheating. All instruments were piloted in two out-of-sample schools before being administered to the study sample.[7]

The baseline survey was administered in July–September 2022, prior to the start of the program. It measured academic achievement (mathematics and science), fluid intelligence, emotional intelligence, experimentation and exploration behavior, risk preferences, and personality traits. The first program year ran during the 2022–23 academic year, followed by a midline survey in February–March 2023. The second program year ran during 2023–24, with the endline survey in March 2024. The endline battery included modified versions of the baseline instruments—adapted to prevent familiarity effects—and added measures of creativity, grit, and problem solving. The endline also included a team-level ideas competition scored by independent evaluators blind to treatment status. Section IV describes each instrument in detail. Finally, government-administered Class X board examinations in March 2025—one full year after the program ended—provide a high-stakes academic outcome free of experimenter demand effects.

Welfare residential schools in Telangana serve students from marginalized communities and experience substantial student turnover: transfers between schools, temporary dropouts, and day-of absences meant that approximately 20% of enrolled students were absent during any given survey wave. Of the 5,642 students on the Class 8 enrollment roster, 4,029 (71.4%) completed the baseline assessment, 4,816 were present at endline, and 4,388 could be matched across waves—yielding a 77.8% baseline-to-endline retention rate.[8] Critically, endline response rates do not differ meaningfully by treatment status: 79.8% for treatment and

[6]The study is registered at the AEA RCT Registry (AEARCTR-0009696).

[7]The tablet-based assessment protocol draws on Singh (2024), who demonstrate that tablet-based assessments reduce cheating and improve measurement validity relative to paper-based formats. Appendix B provides further detail on the field protocol, including enumerator training, device logistics, and data quality procedures.

[8]Cross-wave matching relies on a unique student identifier assigned at enrollment. The gap between the 4,816 students surveyed at endline and the 4,388 matched to the randomization roster reflects students who transferred into sample schools after randomization and were therefore not in the original frame.

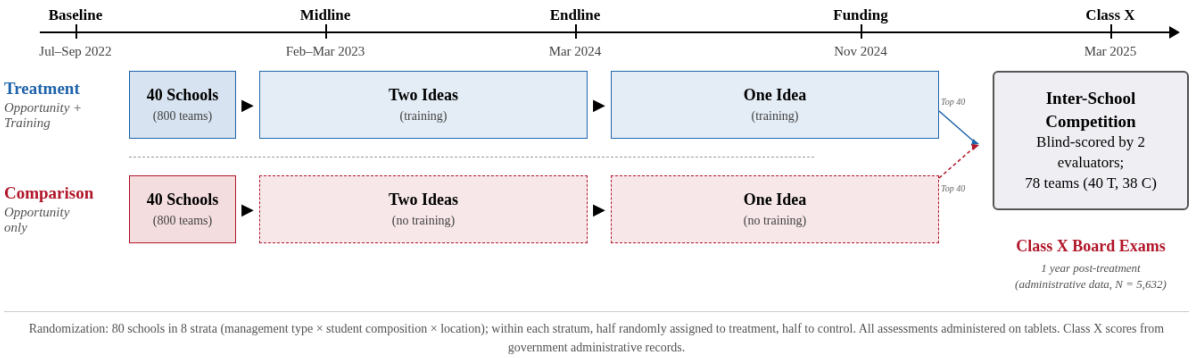


FIGURE 2. RESEARCH DESIGN

Notes: Eighty schools stratified into 8 strata and randomly assigned to treatment or control within each stratum (see text for details). Treatment schools received the two-year innovation curriculum; control schools received only the invitation to the ideas competition. All assessments were administered on tablets. Class X board examination scores come from government administrative records.

75.7% for control, a gap of 4.1 percentage points that is not statistically significant ($p > 0.10$).

To verify that attrition does not compromise the experimental comparison, Appendix Table B.1 reports three checks. First, a joint F-test of treatment-by-covariate interactions on endline attrition fails to reject the null of no differential selection ($F = 1.53$, $p = 0.149$). Second, Lee (2009) trimming bounds preserve the sign and statistical significance of the main results. Third—and most importantly—government-administered Class X board examination scores are available for 5,632 of 5,642 students (99.8%) through administrative records, providing a nearly attrition-free benchmark. Regressing Class X scores on an endline-attrition indicator confirms that students who missed the endline perform indistinguishably from those who attended (mathematics: $\beta = 0.036$ SD, $p = 0.58$; science: $\beta = -0.026$ SD, $p = 0.57$). Because virtually all students sat for the board exams regardless of whether they were present for the research surveys, Class X results confirm the endline findings on academic outcomes in a sample that is not subject to the survey-based attrition documented above.

Table 1 reports summary statistics and tests for baseline balance. Panel A presents student demographics: age, a count of eight household durable assets (air conditioner, car, refrigerator, laptop, scooter, washing machine, and whether each parent owns a mobile phone), and computer literacy.[9] Panel B presents baseline cognitive and non-cognitive measures for the 4,029 baseline respondents. For interpretability, Panel B reports test scores as percent correct rather than as the IRT theta scores used in the regressions: mathematics (10 items, mean 42%), science (15 items, mean 53%), ICAR matrix reasoning (10 items, mean 59%), and the Reading the Mind in the Eyes Test (17 items, mean 53%). Panel B also reports the exploration index from the baseline lemonade-stand game (first principal component of four search behaviors: unique recipes tried, recipe entropy, input changes between rounds, and exploration rate), the Big Five Personality Index (mean of five raw subscale scores with neuroticism reverse-scored), and mean risk preferences from the Gneezy and Potters (1997) investment task (fraction of endowment bet across nine incentivized rounds). Panel C reports school-level infrastructure characteristics from UDISE administrative records ($N = 80$). None of the 20 individual comparisons is statistically significant at conventional levels, and a joint F-test fails to reject the null of no systematic imbalance ($F = 0.87$, $p = 0.561$).

[9]The household asset count in Table 1 reports a simple sum of eight binary indicators for interpretability. In the regressions, the first principal component of the same eight items is used as a continuous socioeconomic status index (PC1 explains 29.8% of variation; durable goods—air conditioner, car, laptop, washing machine—load most heavily). For the 1,022 students (18.1%) missing the SES survey items, the index is imputed at the school mean and a missing-data indicator is included in all regressions.

TABLE 1—SUMMARY STATISTICS AND BASELINE BALANCE

	Control Mean	Treatment Mean	Difference	(SE)	N
Panel A: Student Demographics					
Age (years)	13.35	13.29	−0.061	(0.040)	5,430
Household Assets (0–8)	3.42	3.53	0.115	(0.074)	4,620
Computer Skills (0–1)	0.38	0.40	0.022	(0.029)	4,620
Panel B: Baseline Academic and Higher-Order Skills					
Math Score (%)	42.90	41.92	−0.579	(1.757)	4,029
Science Score (%)	53.35	53.38	0.355	(1.082)	4,029
Fluid Intelligence (%)	59.29	59.43	0.226	(1.480)	4,029
Emotional Intelligence (%)	53.35	52.56	−0.554	(0.871)	4,029
Exploration Index	0.02	−0.02	−0.030	(0.075)	4,029
Big Five Personality Index	26.46	26.23	−0.182	(0.194)	4,029
Risk Preference (% endowment)	62.71	62.87	−0.066	(0.942)	4,025
Panel C: School Characteristics					
Rented Building	0.25	0.38	0.125	(0.105)	80
Has Library	0.50	0.50	−0.000	(0.111)	80
Has Playground	0.12	0.15	0.025	(0.078)	80
English Medium	0.82	0.85	0.025	(0.082)	80
Has Ramp	0.72	0.68	−0.050	(0.104)	80
Has Computer	0.68	0.65	−0.025	(0.106)	80
Boys' Toilets	6.17	6.80	0.625	(2.101)	80
Girls' Toilets	9.25	8.30	−0.950	(2.459)	80
Degraded Boundary Wall	0.07	0.07	0.000	(0.060)	80
No Boundary Wall	0.17	0.17	0.000	(0.084)	80
Joint F-test (p-value)		$F = 0.87$,	$p = 0.561$		

Notes: This table reports baseline means by treatment status and tests for balance across treatment arms. Panel A reports student demographics. Age is computed from date of birth collected across three survey waves; N varies by variable due to item-level missingness (212 students have missing date of birth). Household Assets is a count of 8 household durables (air conditioner, car, refrigerator, laptop, father mobile, mother mobile, scooter, washing machine). Computer Skills is a binary indicator equal to 1 if anyone in the student's household knows how to use a computer (from midline survey). Panel B reports baseline cognitive and non-cognitive measures for the 4,029 students (71.4%) who completed the baseline assessment. Mathematics and Science are percent correct on 10- and 15-item tests, respectively. Fluid Intelligence is percent correct on ICAR matrix reasoning items (10 items; Condon and Revelle, 2014). Emotional Intelligence is percent correct on the Reading the Mind in the Eyes Test (17 items; Baron-Cohen et al., 2001). Exploration Index is the first principal component of four search behaviors (unique recipes tried, recipe entropy, input changes between rounds, exploration rate) from a baseline Lemonade Stand exploration task (adapted from Ederer and Manso, 2013). Big Five Personality Index is the mean of five subscale scores (Agreeableness, Conscientiousness, Extraversion, Openness, and reverse-scored Neuroticism). Risk Preference is the mean fraction of endowment bet across 9 rounds of an incentivized lottery task (Gneezy and Potters, 1997). Panel C reports school-level infrastructure characteristics ($N = 80$). Differences (column 3) are estimated from a regression of each variable on a treatment indicator with strata fixed effects (tribal × girls-only × rural); standard errors (column 4) are clustered at the school level. The joint F-test regresses the treatment indicator on all Panel A and B variables with strata fixed effects, clustered at the school level.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

IV. Measurement

The curriculum targets multiple dimensions of human capital—exploration, fluid intelligence, emotional intelligence, problem solving, persistence, creativity—through a structured cycle of activities that repeat in every project block (Table A.2 summarizes which FLIMP stages target each measured skill). Measuring

whether these skills actually develop requires outcomes that are specific to each targeted skill and that capture revealed behavior rather than self-reports. I use five families of outcome measures. All five families of outcome measures were piloted with out-of-sample students before deployment. Two additional features of the measurement battery are worth noting. First, every behavioral outcome is collected on the same day, on the same tablet, with module order randomized across students—so differences between treatment and control groups cannot be attributed to differential survey fatigue or ordering effects. Second, the measures span a deliberate range of ecological validity: the exploration task and Sudoku puzzles are individual, incentivized, and tightly controlled; the creativity tasks are individual but open-ended; and the ideas competition is team-based, unstructured, and scored by external evaluators blind to treatment status. If treatment effects appear consistently across this range, the finding is unlikely to be an artifact of any single measurement approach. All individual-level outcomes are standardized to mean zero and standard deviation one in the control group; team-level outcomes follow the same convention.

A. Exploration Task

A central claim of this paper is that the FLIMP curriculum builds students' skill in iterative experimentation—generating candidate solutions, testing them against feedback, and refining. To measure this skill, I need a task that captures how students navigate the tension between exploiting a known option and exploring unfamiliar alternatives (March, 1991). Standard survey instruments cannot do this: they record stated preferences, not revealed search behavior. I therefore adapt the multi-armed bandit paradigm from Ederer and Manso (2013) into a lab-in-the-field game that creates a controlled environment in which students must discover profitable configurations through tinkering. The game is designed so that a safe default option is available, the surrounding design space is deliberately sparse, and exceeding the default requires the student to tolerate failure and systematically vary her inputs. Whether a student breaks away from the demonstrated default in round 1—by choosing a stand other than Fresh Juice—provides an immediate signal of willingness to explore an unknown strategy. And because previous-round choices carry forward from round 2 onward, a student who does not actively modify her inputs simply repeats the same configuration, earning the same profit indefinitely. The game thus rewards precisely the kind of iterative experimentation the curriculum is designed to build.

Each student plays individually on an Android tablet. She is told to imagine starting a street food business. A previous owner, Prasuna, is shown to have operated a Fresh Juice stand at the market with 5 menu items, 9 working hours, and a selling price of Rs. 19, earning Rs. 30 per round. In round 1, the student sees Prasuna's configuration as a reference but must make her own selections across five decisions: which of three stands to operate (Fresh Juice, Pani Puri, or Ice Cream), where to locate it (Market, Homes, or School), and how many working hours (6–15), menu items (3–12), and what selling price (Rs. 11–20) to set. Profit is determined by

$$\pi_s(\mathbf{c}) = \max\Big\{0,\ V_s - \lambda_s \sum_d |c_d - c^*_{s,d}|\Big\}, \tag{1}$$

where V_s is the stand's maximum profit, λ_s the per-unit penalty for deviating from the stand-specific optimum, and $c^*_{s,d}$ the optimal input values, which the student does not know. The three stands differ in their risk–return profiles: Fresh Juice offers the lowest ceiling ($V = 120$) but the most forgiving penalty ($\lambda = 10$), Pani Puri a moderate ceiling ($V = 180$) with a steeper penalty ($\lambda = 20$), and Ice Cream the highest ceiling ($V = 250$) but the most severe penalty ($\lambda = 30$). The full payoff landscape spans 9,000 possible combinations, of which 47.7 percent yield zero profit. Payoffs are deterministic and feedback is immediate, so any differences in search behavior reflect differences in willingness and ability to tinker—not noise. The student's stated goal is to find the highest single-round profit in the class for a prize. Appendix Table C.3 reports the complete payoff structure and game parameters.

I construct three outcome variables from the round-level data, each capturing a distinct facet of the

exploration process. *Strategies tried* counts the number of unique stand–location pairs a student visits across all 12 rounds (range: 1–9, since there are 3 stands $\times$ 3 locations), measuring the breadth of tinkering across the choice space. *Discovery count* records the number of rounds, from round 2 onward, in which a student's profit exceeds her running personal best—the maximum profit earned in all prior rounds (range: 0–8). Each such round represents a moment in which tinkering yielded a genuinely new high, so the measure captures the rate at which exploration translates into value discovery. *Minimum distance from optimal* takes, for each round, the sum of absolute deviations between the student's chosen inputs and the stand-specific optimal values across four within-stand dimensions—location, selling price, menu items, and working hours—and selects the minimum of this quantity over all 12 rounds (range: 0–21; lower is better). Unlike maximum profit, which a student may reach partly by chance, the distance measure captures how closely the student's search navigated toward the true optimum and is therefore more indicative of systematic strategy than luck. All three outcomes are standardized to mean zero and standard deviation one in the control group.

Baseline exploration task. A simpler version of the exploration task was administered at baseline, before the start of the program. In the baseline game, students operated a lemonade stand with five input dimensions over 10 rounds (maximum profit Rs 17). Because the baseline and endline games differ in dimensionality, payoff scale, and difficulty—the endline game was redesigned to be harder and more discriminating—raw scores are not directly comparable across waves. The baseline exploration index (first principal component of four search behaviors: unique recipes tried, recipe entropy, input changes between rounds, and exploration rate) serves as the lagged dependent variable for the endline exploration outcomes and as the baseline control for the creativity outcomes, which were measured only at endline.

B. Problem Solving and Grit

The FLIMP curriculum's Make stage requires students to iteratively build and refine solutions under constraints—diagnosing why a prototype failed, adjusting, and trying again. I need a task that captures their skill in structured problem-solving independent of domain knowledge. I also need to measure whether the curriculum builds persistence: willingness to attempt harder challenges despite the risk of failure. Both are measured simultaneously through a Sudoku-based task adapted from Alan et al. (2019), in which the choice of difficulty reveals grit while performance on the puzzles reveals problem-solving ability.

Each student completes five rounds of 4×4 Sudoku puzzles individually on a tablet, with two minutes per round. Before each round, the student chooses between an Easy and a Hard puzzle. One round is randomly selected at the end for a prize: a ballpoint pen if the student chose Easy and solved it correctly, or a more desirable gel pen if she chose Hard and solved it correctly. This design creates a clean incentive tradeoff—the better reward requires both choosing the harder puzzle and executing well—so the choice reveals the student's willingness to take on difficulty under real stakes. Round 1 includes an embedded randomization from Alan et al. (2019): half the students are forced to attempt the Hard puzzle (no choice offered), while the other half choose freely. This exogenous variation in round-1 exposure to difficulty is used to study whether experiencing a hard challenge early shifts subsequent choices, and it ensures that difficulty choice in rounds 2–5 is measured for all students.

I construct two outcome variables from this task. *Problem-solving performance* is the difficulty-weighted average score across all five rounds. Each round's percentage-correct score is multiplied by 1.5 if the student chose Hard and by 1.0 if she chose Easy, and the five weighted scores are averaged. The weight rewards students who both attempt and succeed at harder puzzles, so the measure captures problem-solving ability jointly with ambition rather than performance on trivially easy tasks. *Grit* is estimated from the five binary easy/hard choices using a two-parameter logistic IRT model, which allows items to differ in both difficulty and discrimination—a student who chooses Hard in a round where most students choose Easy receives more credit than one who chooses Hard in an already-popular round. Both outcomes are standardized to mean zero

and standard deviation one in the control group.

C. Fluid Intelligence

The FLIMP curriculum's Learn stage requires students to analyze unfamiliar problems using tools such as cause–effect diagrams and to identify patterns across different contexts. Whether this practice strengthens abstract reasoning is measured using matrix reasoning items adapted from the International Cognitive Ability Resource (ICAR), a public-domain battery validated against the WAIS-IV ($r = .81$) (Condon and Revelle, 2014). Each item presents a 3×3 grid of geometric patterns with one cell missing; the student selects the correct completion from four options (Appendix Figure C.3). The endline assessment comprises 12 items of increasing difficulty, administered individually on a tablet with randomized item order and scored as binary correct/incorrect. Ability is estimated using a two-parameter IRT model (test reliability = 0.981). Answer options were reduced from the original six to four to improve readability on tablet screens. A 10-item version was administered at baseline; the endline version was expanded and recalibrated to increase discrimination at higher ability levels. The baseline IRT theta serves as the lagged dependent variable. The outcome is standardized to mean zero and standard deviation one in the control group.

D. Emotional Intelligence

The FLIMP curriculum's Find stage asks students to identify problems by observing and interviewing stakeholders, an exercise in perspective-taking and reading others' emotional states. I measure this dimension using the Reading the Mind in the Eyes Test (RMET), adapted from Baron-Cohen et al. (2001). Each item presents a photograph of the eye region of a face; the student selects the emotion that best describes what the person is feeling or thinking from four labeled options presented in both English and Telugu (Appendix Figure C.4). The original RMET was developed and validated in Western populations; I adapted the instrument for the Indian context by replacing stimuli with culturally appropriate photographs and piloting the adapted version with 150 out-of-sample students before deployment. The endline assessment comprises 11 items administered individually on a tablet, scored as binary correct/incorrect, with ability estimated using a two-parameter IRT model (test reliability = 0.933). A longer version was administered at baseline; the endline form was shortened to reduce survey fatigue while preserving the most discriminating items. The baseline IRT theta serves as the lagged dependent variable. The outcome is standardized to mean zero and standard deviation one in the control group.

E. Creativity

The FLIMP curriculum's Ideate stage trains students to brainstorm under constraints, evaluate competing ideas against criteria, and generate original solutions to open-ended problems. Whether this training transfers to novel creative tasks—ones unrelated to the curriculum content—requires measures that separately capture written expression, evaluative reasoning, and divergent thinking. I adapt three tasks from the PISA Creative Thinking assessment framework, each targeting a distinct creative competency, and score all responses using a large language model (LLM) with task-specific rubrics.

The first task, *written expression*, presents the student with an image of a dice, globe, and arrow and asks her to write a creative story about it. Responses are scored on a -1 to 5 scale based on narrative structure and incorporation of visual cues: a score of 5 requires multiple sentences that weave in elements of the image (e.g., travel, direction, the world), while a score of -1 indicates a non-interpretable response or a single word. The second task, *idea evaluation*, asks the student to allocate Rs. 1,000 between two inventions—a torch with two bulbs (a simple improvement) and a wheelchair that folds into crutches (a more innovative, multi-functional design)—and to justify her allocation in writing. Responses are scored 0 to 3 based on whether the justification engages with cost, impact, usefulness, or user needs, rather than merely stating a preference. The third task, *idea generation*, asks the student to suggest three improvements to a standard

bicycle. Each idea is scored 0 to 4 on a novelty scale: standard features such as lights or a bell receive a 2, common improvements such as an electric motor or gears receive a 3, and innovative concepts such as solar panels or GPS navigation receive a 4. The task total ranges from 0 to 12.

All three tasks elicit open-ended text responses in English or Telugu-Roman script. Each response is first grammar-corrected and translated to English by the LLM, then graded against the task-specific rubric. To validate the LLM scoring, a research assistant independently scored a random subsample of responses blind to both the LLM's scores and treatment status, applying the same task-specific rubrics. Agreement between the blind human scores and the LLM exceeded 98 percent across all three tasks, with discrepancies confined to a small number of unintelligible or blank responses.[10] Because no baseline creativity measure was collected, the baseline control for all three outcomes is the innovation theta from the baseline lemonade-stand game—the closest available pre-treatment construct. All outcomes are standardized to mean zero and standard deviation one in the control group.

F. Ideas Competition

The individual measures above capture component skills in a controlled setting. To compare innovative output across treatment and control groups, both arms need comparable incentives to produce ideas. At the beginning of each academic year, all 80 schools—treatment and control alike—received a formal letter announcing an end-of-year inter-school innovation challenge (the "Shark Tank Challenge"), emphasizing that teams would be evaluated on how beneficial their idea was for its intended users. Teachers in control schools conducted regular check-ins to ensure their teams were working toward the same deadlines as treatment teams. The treatment–control contrast is therefore structured training (approximately 45 contact hours of FLIMP curriculum) versus unstructured opportunity (same competitive incentive and deadlines, no pedagogical support).

In Year 1, the challenge culminated in a showcase event, but field data collection encountered quality problems: surveyors were unable to transcribe student responses reliably, in part because many students expressed their ideas more fluently in Telugu than in English.[11] The Ideas Competition reported below uses Year 2 submissions only, collected through a redesigned protocol. Teams of three to five students (typically four) work together to develop an idea addressing a real community problem. On the endline survey day, an enumerator visits each school and administers the submission survey using their phone. The team selects a problem sector (e.g., environment, health, crime, agriculture), describes the problem and proposed solution in their own words, identifies target users, explains why the solution improves on existing alternatives, and lists the materials required to build it. The enumerator uploads photographs of the team's labeled drawing and records an audio pitch in which every team member participates. The survey also captures process measures: how many ideas the team considered before settling on one, whether anyone outside the team helped, and how many hours the team spent developing the idea.

I develop the Innovation-S scale, a seven-item instrument measuring two dimensions of innovation quality: novelty (four items: offers a new solution to a problem; improves upon an existing solution; combines existing solutions; transplants a solution from one context to another) and value-added (three items: solves a problem affecting many people; improves the lives of those who will use it; can be easily accessed by those who need it). Each item is rated on a 1–5 scale. I first validate the scale on a sample of 40 adult innovators—entrepreneurs, product managers, and researchers—who rate their own innovations (Appendix Table C.1, Panel A).[12]

[10]The RA subsequently reviewed the LLM's reasoning on all remaining records as an additional quality check; no systematic disagreements were identified beyond the garbage entries flagged in the blind subsample.

[11]Year 1 showcase data are not analyzed in this paper. Transcription failure rates were balanced across treatment and control arms, indicating a mechanical protocol failure rather than differential data loss. The submission protocol was redesigned for Year 2 to use photographs of hand-drawn designs and audio recordings rather than surveyor transcription.

[12]Cronbach's α in the adult validation sample is 0.485, below conventional thresholds. This likely reflects the small sample size

Two evaluators then independently score each student submission blind to treatment status using this scale, along with a separate feasibility item and a hypothetical funding allocation (Rs. 0–1,000). To correct for systematic differences in rater severity—one evaluator scored consistently higher than the other on all dimensions—I convert each evaluator's raw scores to within-evaluator z-scores before averaging, the standard correction in Generalizability Theory (Brennan, 2001). The primary outcome, the *innovation index*, is the equal-dimension-weighted average of the novelty composite and the value-added composite, each given 50 percent weight. In the student sample, Cronbach's α for the seven items is 0.95 and the first eigenvalue captures 77 percent of the variance (Appendix Table C.1, Panel B). Feasibility and funding are treated as separate secondary outcomes. The index is standardized to mean zero and standard deviation one in the control group.

External validation: funding competition. As a further test of innovation quality, a subsample of top-performing teams—selected in equal numbers from treatment and control arms—competed in a live funding competition judged by industry professionals. Nine jury panels, comprising 13 unique jurors, evaluated team pitches; each juror independently allocated between Rs 0 and Rs 10,000 to each team based on innovation quality ($N = 78$ teams: 40 treatment, 38 control). Because jurors differed systematically in leniency and five jurors served on two panels, treatment effects are estimated using a mixed-effects model with juror random intercepts rather than panel-level demeaning. With only 78 teams across 29 school clusters, this competition was never powered for standard hypothesis testing; the results are presented as descriptive evidence.

G. Academic Outcomes

A central question for any curriculum that reallocates instructional time is how it affects academic learning. I measure academic performance at two horizons: an endline assessment administered on the same day as the other individual measures, and the Class X board examinations taken one year later.

The endline assessment includes 22 mathematics items and 20 science items, administered individually on a tablet. Both are scored using a two-parameter logistic IRT model and standardized to mean zero and standard deviation one in the control group. The baseline controls are the corresponding baseline IRT thetas for math and science.

The Class X board examinations are high-stakes state graduation exams administered by the Telangana government in March 2025, one year after the end of the intervention. Because these are administrative records rather than survey data, coverage is near-complete: 5,632 of 5,642 students have valid scores. The exams cover six subjects—mathematics, science, social studies, first language, second language, and third language—each graded on an external proctored exam worth 80 marks. I use external marks only, excluding school-graded internal assessments, to avoid school-level grading noise. All subject scores and the aggregate total are standardized to mean zero and standard deviation one in the control group.

A well-documented limitation of administrative assessments in India is that they systematically compress achievement at the top of the distribution (Singh, 2020). The Class X board exams exhibit this pattern acutely. Among control-group students, board exam scores are left-skewed and cluster near the maximum, so that students who appear identical on the board exam span a wide range of true ability as measured by the IRT assessment. Appendix Figures C.5 and C.6 document this compression in detail using control-group students only: the correlation between the IRT theta and Class X raw score is only $r = 0.22$ for mathematics and $r = 0.31$ for science, and the conditional variance of IRT ability within board exam score deciles rises sharply at the top. This ceiling compression has a direct consequence for interpretation: the board exams have substantially less power than the IRT measure to detect treatment effects at the upper end of the ability

($N = 40$) and restricted variance among experienced innovators who rate their own work highly on most dimensions. In the student sample, where there is substantially more variation in innovation quality, $\alpha = 0.95$ and the first eigenvalue captures 77 percent of item variance, indicating strong unidimensionality.

distribution, and any null results on board exams should be read alongside the IRT estimates.

V. Results

Table 2 reports intent-to-treat estimates for each outcome family. I estimate equation (2) separately for each outcome, with randomization-inference p-values reported alongside conventional standard errors.

A. Estimation Strategy

Intent-to-treat effects are estimated using the following specification:

$$y_{ij}^{E} = \alpha + \beta\, T_j + \gamma y_{ij}^{B} + \delta_k + X_{ij}'\theta + \varepsilon_{ij}, \tag{2}$$

where y_{ij}^{E} is the endline outcome for student i in school j, T_j is an indicator for treatment assignment, y_{ij}^{B} is the baseline analog of the outcome (where available), δ_k denotes strata fixed effects for the eight cells formed by the interaction of school management type (tribal welfare vs. social welfare), student gender (boys' vs. girls' schools), and school location (rural vs. urban), and X_{ij} is a vector of baseline individual controls (age, household assets index, computer literacy) and nine school infrastructure indicators.[13] Where a baseline analog is available, missing values are imputed at the control-group mean and a missing-data indicator is included. Standard errors are clustered at the school level, the unit of randomization, throughout. With 80 school clusters, asymptotic cluster-robust inference may over-reject in finite samples; I therefore report randomization inference (RI) p-values alongside conventional ones in all tables. The coefficient β is the intent-to-treat effect. Because all students in treatment schools were exposed to the curriculum—these are residential schools with near-universal attendance—the ITT closely approximates the average treatment effect.[14]

All individual-level outcomes are standardized to control-group standard deviations. Team-level outcomes (Ideas Competition) are standardized to the graded control-group standard deviation. The primary individual-level analysis sample comprises $N \approx 4{,}388$ students across 80 schools (40 treatment, 40 control). Attrition is balanced across treatment and control arms (Table B.1).

Pre-registration. This study is registered at the AEA RCT Registry (AEARCTR-0009696). All primary outcomes reported below—the exploration task measures, creativity assessments, cognitive skill tests, innovation quality scores, willingness to fund, and academic outcomes—were pre-specified in the registry or in the detailed assessment instruments uploaded in January 2024. The Creative Problem-Solving (CPS) composite reported in Table 2 was not pre-registered as a named index; it is an equal-weighted average of all pre-registered individual skill outcomes, constructed for ease of interpretation. Table D.1 in the appendix shows that all results are stable across alternative control specifications; Table D.2 demonstrates robustness to alternative index constructions including Anderson inverse-covariance weighting and inclusion of all available skill measures.[15]

[13] School infrastructure controls: rented premises, boys' toilets, girls' toilets, library, playground, medical facilities, computer access, wall condition (degraded, absent).

[14] Implementation tracking data confirm that 43 of 45 monitored treatment schools completed all curriculum sessions (96% completion rate). The two remaining schools completed 3 and 6 of 7 sessions respectively.

[15] Four pre-registered outcomes are not included in this paper. Big Five personality traits and the risk-taking task (Gneezy and Potters, 1997) were administered at midline but dropped from the endline battery—the risk-taking task showed no variation and Big Five is self-reported—to reduce test fatigue and accommodate new performance-based assessments (creativity, problem solving, grit). The Innovators DNA survey was never collected; it was replaced by the exploration task before baseline. Social network data are reserved for a companion paper examining team composition effects, as teams were separately randomized on social skills at baseline in both treatment and control schools. Appendix F provides a complete mapping of pre-registered outcomes to results.

Multiple comparisons. This paper reports treatment effects on distinct pre-registered outcome families—exploration behavior, innovation quality, individual cognitive and creative skills, and academic performance—each corresponding to a separate theoretical channel. Because these families test substantively different hypotheses rather than multiple draws from a single test, I follow the approach of reporting pre-registered outcomes individually with randomization inference p-values throughout, rather than applying a family-wise error rate correction across unrelated domains.[16]

B. Main Effects on Creative Problem-Solving

Table 2 presents the headline results. The innovation curriculum increases the CPS composite by 0.120 standard deviations (SE = 0.031, $p < 0.001$, RI $p = 0.003$). This composite is the equal-weighted mean of three domain z-scores: Creativity (written expression, idea evaluation, and idea generation), Analytical Reasoning (fluid intelligence and problem solving), and Exploration (discovery count, strategies tried, and proximity to optimal). The treatment effect on the academic composite—the mean of endline mathematics and science IRT theta scores—is 0.011 standard deviations and statistically indistinguishable from zero. The curriculum replaced one 90-minute homework period per week with innovation activities; despite this reallocation of instructional time, there is no evidence of academic harm.

Figure 3 summarizes treatment effects across all outcomes. The remaining subsections unpack each outcome family in detail.

TABLE 2—MAIN TREATMENT EFFECTS

	Effect (σ)	(SE)	[RI p]	N
Creative Problem-Solving (CPS)	0.120***	(0.032)	[0.003]	4,388
Academic	0.011	(0.056)	[0.868]	4,388

Notes: Each row reports the OLS estimate of the treatment effect on a pre-specified composite index. Creative Problem-Solving (CPS) is the equal-weighted mean of three domain z-scores: (a) Creativity (written expression, idea evaluation, idea generation), (b) Analytical Reasoning (fluid intelligence, problem solving), and (c) Exploration (discovery count, strategies tried, proximity to optimal). Academic is the mean of endline mathematics and science IRT theta scores. All component outcomes are standardized to the control-group mean and SD before averaging. Each domain receives equal weight (1/3 for CPS, 1/2 for Academic). All specifications include strata fixed effects (tribal × girls-only × rural), baseline individual controls (age, household assets index, computer literacy), 9 school infrastructure controls, and baseline analogs of the component outcomes where available; missing baseline values are imputed at the control mean with a missing indicator. Standard errors clustered at the school level (80 clusters) in parentheses. Randomization inference p-values from 999 within-strata school-level permutations in square brackets. Component-level results are reported in Tables 3–5. Alternative specifications are shown in Table D1; robustness to alternative CPS definitions in Table D2. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

C. Exploration Task: Process Evidence

The exploration task (Section IV) yields three cumulative outcomes: Discovery Count (running personal best profit through round r), Strategies Tried (unique stand-location combinations explored), and Proximity to Optimal (closest Manhattan distance to any stand's optimal input vector).

Panel A of Table 3 reports treatment effects measured through round 12. Treated students achieve a higher running personal best: Discovery Count increases by 0.094 SD ($p = 0.012$, RI $p = 0.030$). They search more broadly, trying 0.077 SD more unique stand-location pairs (Strategies Tried; $p = 0.073$, RI $p = 0.136$).

[16]The CPS composite serves as a summary index in the spirit of Anderson (2008), aggregating across individual skill measures to reduce the dimensionality of inference. Table D.2 shows that this composite is robust to Anderson inverse-covariance weighting, which optimally accounts for the covariance structure across outcomes. The individual-outcome results in Tables 4 and 3 are presented to decompose this aggregate effect, not as independent hypothesis tests. Randomization inference p-values, reported for every estimate, provide exact finite-sample inference without distributional assumptions.

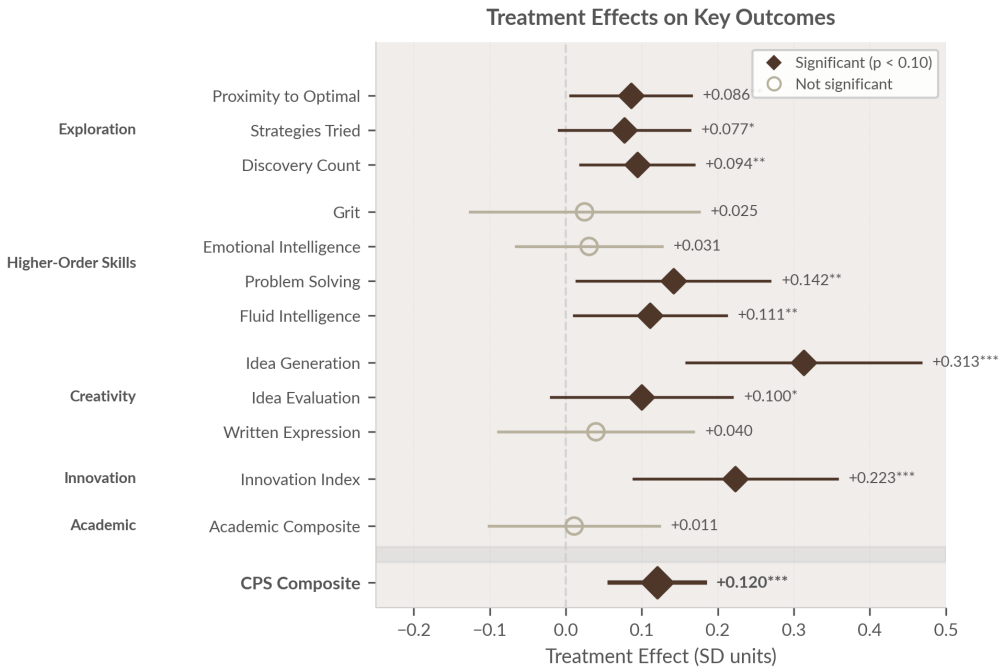


FIGURE 3. TREATMENT EFFECTS ON CREATIVE PROBLEM-SOLVING, INNOVATION, AND ACADEMIC OUTCOMES

Notes: Filled diamonds indicate significance at the 10% level; open circles indicate non-significance. Horizontal lines show 95% confidence intervals. All specifications include strata fixed effects, baseline individual controls, 9 school infrastructure controls, and baseline analogs where available. Standard errors clustered at the school level (80 clusters). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Their search is also directed rather than random: Proximity to Optimal improves by 0.086 SD ($p = 0.031$, RI $p = 0.069$).

Appendix Figure C.1 and Appendix Table C.4 trace these dynamics round by round. The treatment–control gap builds progressively: Proximity to Optimal diverges first (significant from round 5), followed by Strategies Tried (rounds 6–7), and finally the running personal best (rounds 9–11). This sequencing is consistent with a compounding search advantage—treated students try more combinations each round, raising the probability of beating their personal best, and these small advantages accumulate over the 12-round horizon.

D. Innovation Outcomes: Ideas Competition

The exploration task provides controlled evidence on search behavior in a laboratory setting. Does this translate into real-world innovative output? Panel B of Table 3 reports treatment effects on the year-end Ideas Competition (Section IV), in which teams of 3–5 students developed and submitted innovation proposals scored by external evaluators blind to treatment status ($N = 1{,}283$ scored teams).

The composite Innovation Index increases by 0.223 SD ($p = 0.001$, RI $p = 0.005$). Among individual components, the largest effect is on Novelty (0.263 SD, $p < 0.001$), followed by Value-Add (0.190 SD, $p = 0.007$) and Willingness to Fund (0.178 SD, $p = 0.023$). Feasibility shows a smaller, marginally significant effect (0.105 SD, $p = 0.081$). Taken together, the exploration task and the Ideas Competition provide convergent evidence from two distinct measurement approaches—an incentivized individual-level laboratory task and a team-level field innovation scored by independent evaluators—that the curriculum builds transferable innovation skills.

TABLE 3—INNOVATION OUTCOMES

	Effect (σ)	(SE)	[RI p]	N
Panel A: Exploration Behavior				
Discovery Count	0.094**	(0.037)	[0.030]	4,388
Strategies Tried	0.077*	(0.043)	[0.136]	4,388
Proximity to Optimal	−0.086**	(0.040)	[0.069]	4,388
Panel B: Idea Quality				
Innovation Index	0.223***	(0.067)	[0.005]	1,283
Novelty	0.263***	(0.071)	[0.004]	1,189
Value-Add	0.190***	(0.070)	[0.033]	1,189
Feasibility	0.105*	(0.060)	[0.175]	1,189
Willingness to Fund	0.178**	(0.078)	[0.062]	1,189

Notes: This table reports OLS estimates of equation (1). The dependent variable in each row is regressed on a treatment indicator. Panel A uses the individual-level endline sample across 80 schools (40 treatment, 40 control). Outcomes are from a 12-round exploration task adapted from Ederer and Manso (2013), in which students choose inputs (stand type, location, hours, menu, price) to maximize profit. Discovery Count is the number of unique positive-profit strategies found; Strategies Tried is the number of unique (stand, location) combinations explored; Proximity to Optimal is the negative of the minimum Manhattan distance from any optimal input vector. All outcomes are standardized to the control-group mean and SD. All specifications include strata fixed effects (tribal × girls-only × rural), baseline individual controls (age, household assets index, computer literacy), 9 school infrastructure controls, and the baseline analog of the outcome where available; missing baseline values are imputed at the control mean with a missing indicator. Panel B uses scored submissions from the ideas competition. Innovation Index is a 50/50 weighted average of the Novelty (4 items) and Value-Add (3 items) sub-indices; all item scores are within-evaluator z-scored to remove rater severity effects, then control-group standardized. The Innovation Index sample ($N = 1{,}283$) includes 94 teams whose submissions were recovered from the pipeline after photo validation; their index is mapped from aggregate innovation scores using a linear fit on the 1,189 evaluator-graded teams ($r = 0.99$). Sub-component estimates use only the 1,189 evaluator-graded teams with item-level data. Panel B specification includes strata fixed effects and school infrastructure controls (no individual-level controls). Standard errors clustered at the school level in parentheses. Randomization inference p-values from 999 within-strata school-level permutations in square brackets. Stars are from asymptotic (OLS) inference.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

E. Descriptive Evidence: Innovation Funding Competition

As a final test of innovation quality, a subsample of top-performing teams was selected—in equal numbers by design—from approximately 800 teams per arm to compete in a live funding competition (Section IV). Panels of industry professionals evaluated pitches and allocated funding from individual budgets of Rs 0–10,000 per juror ($N = 78$ teams: 40 treatment, 38 control; 13 jurors across 9 panels).

Figure 4 presents the results. Panel A shows CDFs of total jury funding (sum of both jurors' allocations) for treatment and control teams. A mixed-effects model with juror random intercepts estimates a funding gap of Rs 542 per juror ($p = 0.260$; Cohen's $d = 0.18$). Panel B decomposes this into quantile treatment gaps: treatment teams lose at the bottom of the distribution (Q5–Q15), dominate in the middle (Q25–Q50: Rs +1,000 to +3,300), and pull ahead again at the top (Q80–Q95: Rs +2,500 to +4,550). Treatment teams are overrepresented among the highest-funded: 8 of the 10 teams receiving ≥Rs 14,000 are from treatment schools (Fisher $p = 0.052$). With only 78 teams, this competition is underpowered for standard hypothesis testing; these results are presented as descriptive evidence consistent with the innovation quality effects documented above.

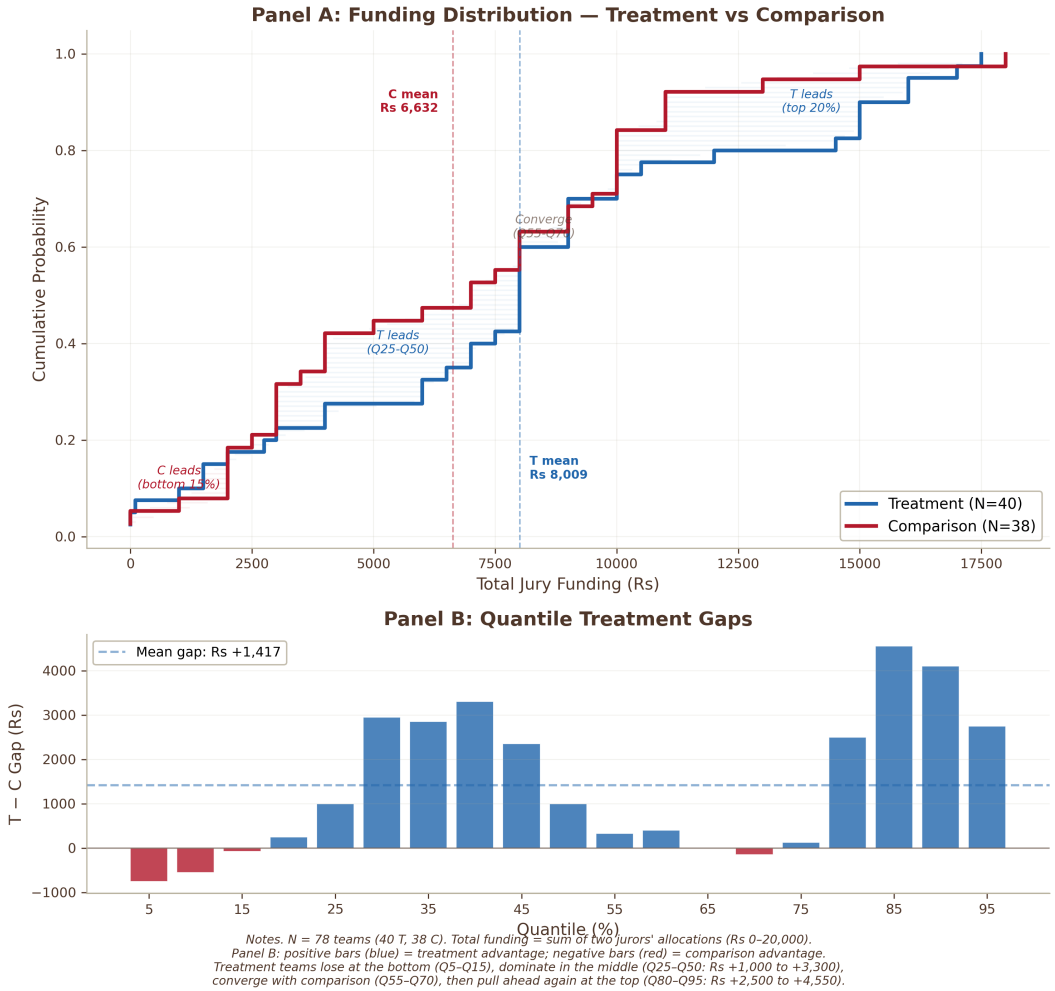


FIGURE 4. INNOVATION FUNDING COMPETITION — CDF AND QUANTILE TREATMENT GAPS

Notes: Panel A: CDFs of total jury funding for treatment and control teams ($N = 78$: 40 treatment, 38 control). Dashed lines indicate group means. Panel B: quantile treatment gaps (treatment − control) at each ventile. Mixed-model estimate: +Rs 542 ($p = 0.260$).

F. Higher-Order Skills

The innovation outcomes demonstrate that treated students explore better and produce higher-quality innovations. Table 4 reports treatment effects on seven individual skill measures grouped into three domains.

Analytical Reasoning. Treated students score 0.111 SD higher on fluid intelligence ($p = 0.028$, RI $p = 0.059$) and 0.142 SD higher on problem solving ($p = 0.028$, RI $p = 0.061$). Fluid intelligence is typically considered difficult to shift through educational interventions; this result is consistent with the hypothesis that sustained practice in structured exploration and problem decomposition can strengthen abstract reasoning.

Creativity. Effects are largest on Idea Generation (0.313 SD, $p < 0.001$, RI $p = 0.002$). Idea Evaluation shows a more modest effect (0.100 SD, $p = 0.096$, RI $p = 0.158$), while Written Expression is not significantly affected (0.040 SD). The pattern is informative: the curriculum emphasizes brainstorming, prototyping, and iterative refinement, precisely the skills captured by idea generation, but does not specifically train written communication.

TABLE 4—HIGHER-ORDER SKILLS

	Effect (σ) (SE)		[RI p]	N
Panel A: Analytical Reasoning				
Fluid Intelligence	0.111**	(0.050)	[0.059]	4,388
Problem Solving	0.142**	(0.064)	[0.061]	4,388
Panel B: Creativity (PISA)				
Written Expression	0.040	(0.065)	[0.601]	4,388
Idea Evaluation	0.100*	(0.060)	[0.158]	4,388
Idea Generation	0.313***	(0.077)	[0.002]	4,388
Panel C: Socio-Emotional				
Emotional Intelligence	0.031	(0.048)	[0.597]	4,388
Grit	0.025	(0.076)	[0.753]	4,388

Notes: This table reports OLS estimates of equation (1). The dependent variable in each row is regressed on a treatment indicator. All regressions use the individual-level endline sample across 80 schools (40 treatment, 40 control). All outcomes are standardized to the control-group mean and SD. All specifications include strata fixed effects (tribal × girls-only × rural), baseline individual controls (age, household assets index, computer literacy), 9 school infrastructure controls, and the baseline analog of the outcome where available; missing baseline values are imputed at the control mean with a missing indicator. Panel A: Fluid Intelligence is measured by ICAR matrix reasoning items (12 items; Condon and Revelle, 2014); Problem Solving by a difficulty-weighted Sudoku task (5 rounds, 4×4 grid; each round the student chooses an Easy or Hard puzzle; hard puzzles receive weight 1.5; Newell and Simon, 1972). Panel B: Creativity is measured by the PISA 2022 Creative Thinking assessment (OECD, 2024). Idea Generation is the number of original ideas produced in a timed task; Idea Evaluation measures the ability to rank ideas by quality; Written Expression is a holistic score for a short essay. Panel C: Emotional Intelligence is measured by the Reading the Mind in the Eyes Test (11 items; Baron-Cohen et al., 2001; Weidmann and Deming, 2021, show RMET predicts team player skills). Grit is measured by a 5-item grit scale embedded in an Easy/Hard choice task (Alan et al., 2019). Standard errors clustered at the school level in parentheses. Randomization inference p-values from 999 within-strata school-level permutations in square brackets. Stars are from asymptotic (OLS) inference.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Socio-Emotional Skills. Treatment effects on Emotional Intelligence (0.031 SD) and Grit (0.025 SD) are small and statistically insignificant. These null effects serve as a useful placebo: they suggest that the positive results on cognitive and creative skills are not driven by Hawthorne effects, increased motivation, or generalized attention from participating in a novel program. The curriculum appears to build specific cognitive skills rather than shifting broad dispositional traits.

G. Academic Outcomes

A key policy question about innovation-focused curricula is how they affect core academic performance. In this setting, the program replaced one weekly homework period (90 minutes) with structured innovation activities. The analysis tests for academic effects—both negative and positive—using two independent data sources.

First, as reported in Table 2, the endline academic composite—the mean of mathematics and science IRT theta scores administered as part of the study's assessment battery—shows a precisely estimated null effect of 0.011 SD (SE = 0.056).[17]

[17]Year 1 midline academic assessments show a transient negative treatment effect on mathematics (−0.14 SD, RI $p = 0.021$) that fully reverses by endline and Class X boards, consistent with a survey context artifact driven by the placement of cognitive items immediately after an engaging creative task. Supplementary Appendix S documents a comprehensive mechanism investigation

TABLE 5—CLASS X BOARD EXAMINATION RESULTS

	Effect (σ) (SE)		[RI p]	N
Mathematics	0.112	(0.147)	[0.533]	5,632
Science	−0.006	(0.119)	[0.978]	5,632
Social Studies	0.023	(0.093)	[0.818]	5,632
First Language	0.149*	(0.088)	[0.161]	5,632
Second Language	0.199*	(0.106)	[0.101]	5,631
Third Language	0.153*	(0.086)	[0.116]	5,631
Total Score	0.135	(0.115)	[0.326]	5,633
Passed	0.006	(0.008)	[0.532]	5,642

Notes: This table reports OLS estimates of equation (1) using administrative Class X board examination records from the Telangana State Board of Secondary Education. The sample covers the near-universe of students in the original randomization. All outcomes except "Passed" are standardized to the control-group mean and SD. "Passed" is a binary indicator (1 if student passed all subjects). All specifications include strata fixed effects (tribal × girls-only × rural), baseline individual controls (age, household assets index, computer literacy), 9 school infrastructure controls, and the baseline analog of the outcome where available; missing baseline values are imputed at the control mean with a missing indicator. Standard errors clustered at the school level in parentheses. Randomization inference p-values from 999 within-strata school-level permutations in square brackets. Stars are from asymptotic (OLS) inference. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Second, administrative data from the Telangana State Board Class X examinations, taken by students approximately one year after the intervention ended. These high-stakes exams are externally administered, externally graded, and determine secondary school completion—they are immune to experimenter demand effects and provide a longer-run test of academic persistence. Table 5 reports effects on eight board exam outcomes for the near-universe of students in the original randomization ($N \approx 5{,}632$). Treatment effects on mathematics (0.112 SD, $p = 0.446$), science (−0.006 SD), and total score (0.135 SD, $p = 0.241$) are all statistically insignificant, with point estimates that are uniformly positive or zero. Language subjects show marginally positive effects: First Language (0.149 SD, $p = 0.091$), Second Language (0.199 SD, $p = 0.060$), and Third Language (0.153 SD, $p = 0.074$). The pass rate is unaffected (0.006, $p = 0.466$). These results indicate that the innovation curriculum does not reduce academic performance. That all three language subjects move in the same direction—while math and science do not—is unlikely to be coincidental, and the pattern is consistent with the curriculum's emphasis on articulation. FLIMP sessions require students to brainstorm aloud, pitch ideas to peers, and justify design choices in writing; these are expressive skills that transfer directly to language examinations. The largest effect appears on Second Language (Telugu, the mother tongue), consistent with students learning to express themselves actively rather than engaging as passive learners. Section B examines this possibility further through a quintile decomposition of the ability distribution.

VI. Heterogeneous Treatment Effects

Who benefits most? I examine heterogeneity along four dimensions: baseline academic ability, gender, school management type (tribal vs. social welfare), and school location (rural vs. urban). All heterogeneity analyses use an interacted specification:

$$y_{ij}^{E} = \beta_0 + \beta_1 T_j + \beta_2 G_{ij} + \beta_3 (T_j \times G_{ij}) + X_{ij}' \theta + \varepsilon_{ij}, \tag{3}$$

testing 12 alternative hypotheses.

where G_{ij} is the group indicator. The heterogeneity dimension is removed from the strata fixed effects and entered directly with its interaction. The tables report β_1 (the treatment effect for the reference group), $\beta_1 + \beta_3$ (the effect for the comparison group), and the p-value on β_3 (the test of differential effects). Table E.1 in the appendix extends the analysis to all individual outcomes not included in the main heterogeneity table.

A. Baseline Academic Ability

The most policy-relevant heterogeneity question is whether the program only benefits students who are already academically strong—in which case it would reinforce existing advantages—or whether it reaches students across the ability distribution. Table 6, Panel A reports treatment effects separately for students below and above the median of baseline academic ability (a combined IRT theta from all 27 baseline mathematics and science items; median computed on the control group).

The CPS composite effect is significant in both groups but significantly larger for above-median students: 0.196 SD versus 0.094 SD ($p_{\text{diff}} = 0.034$). Decomposing this into individual tasks reveals a nuanced pattern. On exploration task outcomes, effects are similar or larger for below-median students: Discovery Count shows a treatment effect of 0.111 SD below the median versus 0.075 SD above ($p_{\text{diff}} = 0.620$). The program teaches students of all ability levels to explore more effectively. By contrast, effects on idea generation and fluid intelligence are concentrated among above-median students: Idea Generation increases by 0.457 SD for above-median versus 0.198 SD for below-median students ($p_{\text{diff}} = 0.012$), and Fluid Intelligence by 0.222 SD versus 0.020 SD ($p_{\text{diff}} = 0.014$). The interaction on the academic composite is also significant ($p_{\text{diff}} = 0.042$), a pattern examined more closely with the quintile decomposition below. Appendix Figure E.1 presents the full set of outcome-by-outcome heterogeneity estimates; across all 13 outcomes, treatment effects are positive for both groups.

B. Who Develops Creative Problem-Solving?

The median split in Table 6 masks heterogeneity within the above- and below-median groups. Figure 5 presents a binned scatterplot of endline CPS against baseline academic ability, showing that the treatment shifts CPS upward across the entire ability distribution, with the treatment–control gap widening at higher baseline ability levels.

Figure 6 decomposes these effects by baseline ability quintile. Panel A shows that the treatment raises CPS in every quintile, with the largest gains at the top (Q5: $+0.258$ SD, $p < 0.001$) and bottom (Q1: $+0.146$ SD, $p < 0.001$) of the distribution. Panel B reveals a striking asymmetry in academic outcomes. For students in the bottom four quintiles, the innovation curriculum produces no detectable change in endline mathematics or science scores — the null effects on academic achievement documented in Section G reflect this group. For students in the top quintile, however, the treatment *raises* academic performance by 0.194 SD ($p = 0.021$), driven primarily by mathematics ($+0.218$ SD, $p = 0.012$). This pattern — null at the bottom, significantly positive at the top — indicates cognitive transfer of the program in mathematics for top-quintile students. The result is robust to alternative specifications (bare through full KLK), alternative cutpoints (top tercile through top quintile), and a formal full-sample interaction test (treatment-by-top-quintile $p = 0.019$ on the composite, $p = 0.003$ on mathematics); the randomization inference p-value for mathematics is 0.046. Appendix Table E.3 documents these checks in detail. Table E.2 in the appendix reports the full set of quintile-specific estimates across all CPS and academic outcomes.

The absence of a corresponding gradient in Class X board examinations is expected given the severe ceiling compression documented in Appendix Figures C.5–C.6: the correlation between IRT ability and board exam scores is just 0.22 in mathematics, substantially limiting the power to detect ability-specific effects on an externally graded assessment that does not discriminate well at the top of the distribution.

To further characterize how the program shifts the joint distribution of academic ability and creative problem-solving, Figure 7 classifies students into four profiles by crossing baseline academic ability (above

TABLE 6—HETEROGENEOUS TREATMENT EFFECTS

	Group 1	(SE)	Group 2	(SE)	p(diff)
Panel A: By Baseline Academic Ability					
	Below Median		Above Median		
Creative Problem-Solving (CPS)	0.094***	(0.035)	0.196***	(0.042)	0.034
Discovery Count	0.111**	(0.046)	0.075	(0.058)	0.620
Proximity to Optimal	−0.090*	(0.046)	−0.081	(0.056)	0.885
Idea Generation	0.198***	(0.071)	0.457***	(0.102)	0.012
Fluid Intelligence	0.020	(0.063)	0.222***	(0.063)	0.014
Problem Solving	0.158**	(0.070)	0.126	(0.083)	0.697
Math & Science	−0.061	(0.055)	0.097	(0.079)	0.042
Panel B: By Gender					
	Boys' Schools		Girls' Schools		
Creative Problem-Solving (CPS)	0.080*	(0.046)	0.190***	(0.045)	0.104
Discovery Count	0.012	(0.042)	0.165***	(0.060)	0.044
Proximity to Optimal	−0.060	(0.053)	−0.109*	(0.055)	0.508
Idea Generation	0.095	(0.087)	0.504***	(0.123)	0.010
Fluid Intelligence	0.062	(0.069)	0.153**	(0.069)	0.339
Problem Solving	0.250**	(0.115)	0.046	(0.076)	0.164
Math & Science	−0.032	(0.081)	0.049	(0.079)	0.486
Panel C: By School Management					
	Scheduled Tribes		Scheduled Castes		
Creative Problem-Solving (CPS)	0.058	(0.043)	0.210***	(0.047)	0.023
Discovery Count	0.128**	(0.055)	0.064	(0.054)	0.434
Proximity to Optimal	−0.066	(0.063)	−0.104*	(0.056)	0.669
Idea Generation	0.150	(0.103)	0.457***	(0.118)	0.059
Fluid Intelligence	−0.004	(0.073)	0.213***	(0.074)	0.050
Problem Solving	0.088	(0.095)	0.189**	(0.094)	0.469
Math & Science	−0.114	(0.070)	0.122	(0.083)	0.032
Panel D: By Location					
	Urban		Rural		
Creative Problem-Solving (CPS)	0.120*	(0.062)	0.149***	(0.038)	0.697
Discovery Count	0.072	(0.062)	0.106**	(0.047)	0.663
Proximity to Optimal	−0.090	(0.069)	−0.083*	(0.049)	0.940
Idea Generation	0.347**	(0.157)	0.294***	(0.086)	0.773
Fluid Intelligence	0.104	(0.093)	0.115*	(0.067)	0.934
Problem Solving	0.017	(0.120)	0.210***	(0.076)	0.185
Math & Science	0.106	(0.108)	−0.041	(0.074)	0.301
N			4,388		

Notes: Each row reports estimates from an interacted regression: $Y = \beta_0 + \beta_1 T + \beta_2 \text{Group} + \beta_3 (T \times \text{Group}) + \mathbf{X}\gamma + \varepsilon$. The Group 1 column reports $\hat{\beta}_1$ (the treatment effect for the reference group); the Group 2 column reports $\hat{\beta}_1 + \hat{\beta}_3$ (the treatment effect for the other group), with standard error from the full variance-covariance matrix. The p(diff) column tests H_0: $\beta_3 = 0$. Outcomes are defined in Tables 3–5. Creative Problem-Solving (CPS) is the equal-weighted mean of three domain composites: Creativity (idea evaluation and idea generation), Analytical Reasoning (fluid intelligence and problem solving), and Exploration (discovery count, strategies tried, proximity to optimal). Math & Science is the mean of endline mathematics and science IRT theta scores. Reference groups: Panel A, below-median baseline academic ability; Panel B, boys' schools; Panel C, Scheduled Tribes (ST) schools; Panel D, urban schools. Panels B–D split on school-level stratification variables; the het dimension is removed from strata fixed effects and entered directly with its interaction. Panel A splits at the control-group median of a combined IRT theta estimated from all 27 baseline math and science items pooled into a single 2PL model. All specifications include strata fixed effects, baseline individual controls, 9 school infrastructure controls, and the baseline analog of the outcome where available. Standard errors clustered at the school level (80 clusters) in parentheses. Table E1 reports heterogeneous effects on the remaining outcomes.

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

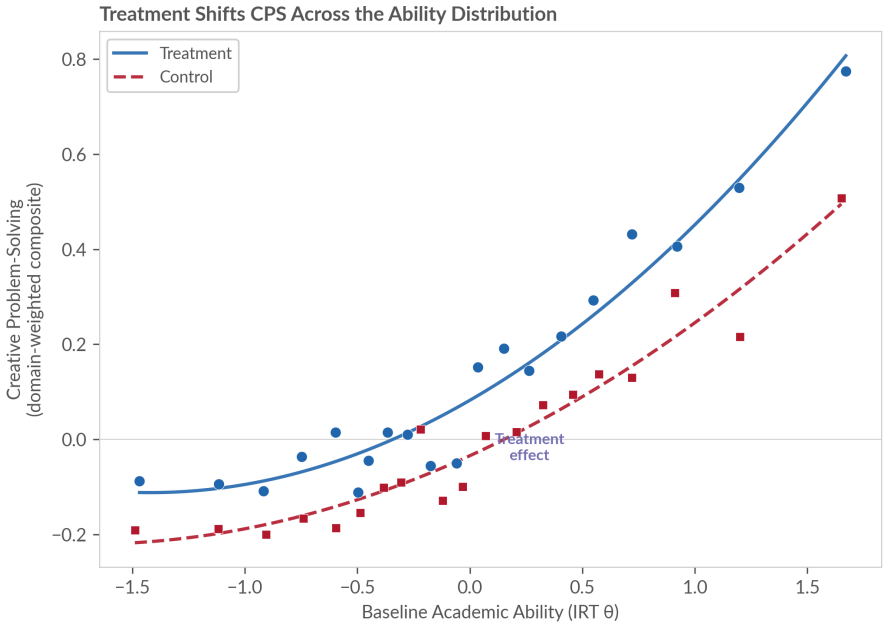


FIGURE 5. TREATMENT SHIFTS CPS ACROSS THE ABILITY DISTRIBUTION

Notes: Binned scatterplot (20 equal-sized bins) of endline Creative Problem-Solving composite against baseline academic ability, separately for treatment and control, with quadratic polynomial fits. The CPS composite is the equal-weighted mean of Creativity, Analytical Reasoning, and Exploration domain z-scores. $N = 4{,}388$.

vs. below the overall median) with endline CPS (above vs. below the *control-group* median). Among students above the baseline academic median ($N = 2{,}194$), treatment increases the share classified as "High Talent"—above median on *both* dimensions—by 10.4 percentage points ($p < 0.001$). These are students for whom the quintile analysis suggests innovation and academic gains reinforce each other. Among students below the baseline academic median ($N = 2{,}194$), treatment increases the share classified as "Emerging CPS"—below median on academics but above median on CPS—by 7.3 percentage points ($p < 0.001$). In a system that selects on academic ability alone, these students' creative potential would remain invisible. This result—which is labeled exploratory, as the profile framework was developed ex post—suggests that innovation curricula can identify and cultivate talent that traditional metrics miss. Appendix Figure E.2 replicates the quintile analysis separately for each CPS sub-component.

C. Gender

Panel B of Table 6 and Appendix Figure E.3 report treatment effects separately for boys' and girls' schools. The CPS composite rises by 0.190 SD in girls' schools versus 0.080 SD in boys' schools ($p_{\text{diff}} = 0.104$). The individual tasks reveal where this gap originates. On innovation outcomes, girls' schools show substantially larger effects than boys' schools: Discovery Count increases by 0.165 SD in girls' schools versus 0.012 SD in boys' schools ($p_{\text{diff}} = 0.044$); Idea Generation increases by 0.504 SD versus 0.095 SD ($p_{\text{diff}} = 0.010$). These differences are statistically significant and economically large. On cognitive outcomes, Fluid Intelligence shows a larger effect for girls (0.153 SD) than boys (0.062 SD), though the difference is not significant ($p_{\text{diff}} = 0.339$). Problem Solving reverses: 0.250 SD for boys versus 0.046 SD for girls ($p_{\text{diff}} = 0.164$). Because all schools in the sample are single-gender, it is not possible to separate the effect of gender from the effect of single-gender classroom dynamics.[18]

[18]The heterogeneity analysis compares boys' schools to girls' schools, not boys to girls within schools. Differences could reflect gender, classroom composition, peer effects, or teacher characteristics that differ systematically across single-gender school types.

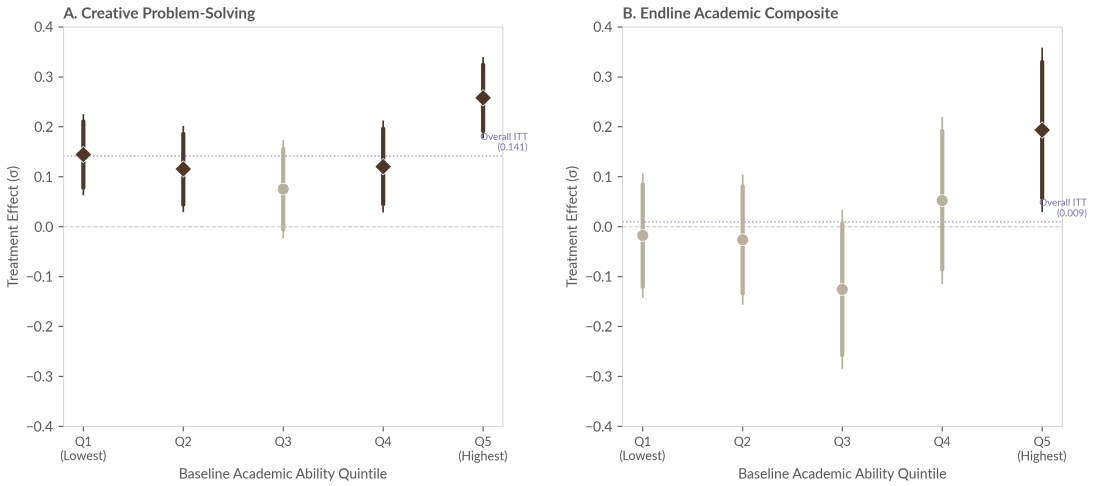


FIGURE 6. TREATMENT EFFECTS ON CPS AND ACADEMIC OUTCOMES BY BASELINE ABILITY QUINTILE

Notes: Each point shows the treatment effect from a separate OLS regression estimated within the indicated baseline ability quintile. Thick bars = 90% CI; thin bars = 95% CI. Filled diamonds indicate $p < 0.10$; open circles indicate $p \geq 0.10$. The dotted line shows the full-sample ITT. Panel A outcome: CPS composite (equal-weighted average of Creativity, Analytical Reasoning, and Exploration domains). Panel B outcome: Academic composite (average of endline Mathematics and Science IRT theta scores, control-group standardized). All regressions include strata FE, baseline individual controls, 9 school infrastructure controls, and baseline academic ability. Standard errors clustered at the school level. $N \approx 4{,}388$.

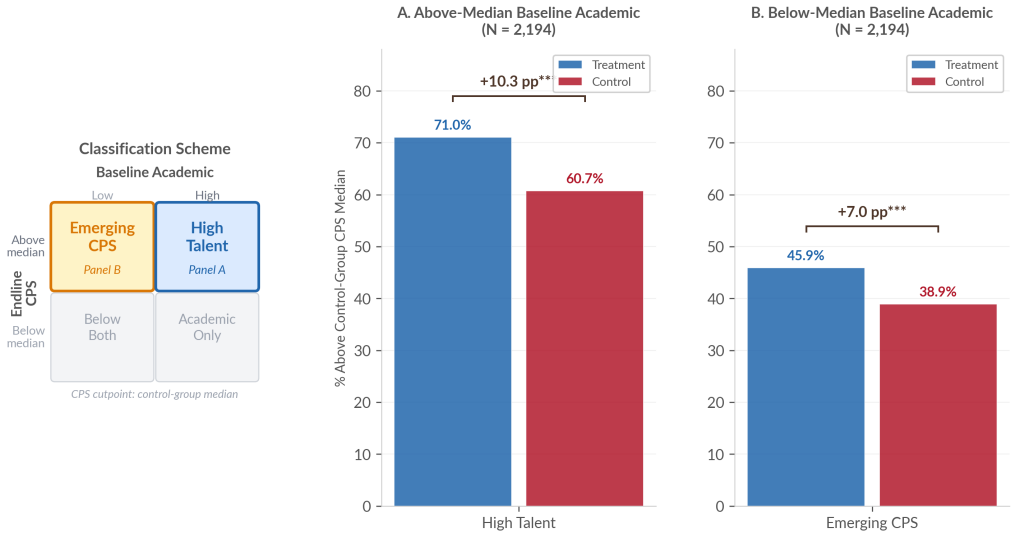


FIGURE 7. WHO DEVELOPS CREATIVE PROBLEM-SOLVING? TREATMENT EFFECTS BY BASELINE ABILITY

Notes: Students classified into four profiles by baseline academic ability (overall median) × endline CPS (control-group median). The CPS composite is the equal-weighted mean of Creativity, Analytical Reasoning, and Exploration domain z-scores. Panel A: above-median baseline academic ($N = 2{,}194$); shows the share classified as "High Talent." Panel B: below-median baseline academic ($N = 2{,}194$); shows the share classified as "Emerging CPS." Treatment effects from LPMs with strata FE, baseline controls, infrastructure controls. SEs clustered at school level.

D. School Management and Location

Panels C and D of Table 6 examine heterogeneity by school management type and location. The CPS composite effect is 0.210 SD in social welfare (SC) schools versus 0.058 SD in tribal welfare (ST) schools

($p_{\text{diff}} = 0.023$). This gap is visible across individual tasks: Idea Generation is 0.457 SD in social welfare versus 0.150 SD in tribal welfare ($p_{\text{diff}} = 0.059$), and Fluid Intelligence is 0.213 SD versus -0.004 SD ($p_{\text{diff}} = 0.050$). SC schools enter with baseline academic ability 0.40 SD above ST schools ($p < 0.001$); given the ability gradient documented in Panel A, this differential likely reflects the concentration of higher-ability students in SC schools rather than a pure institutional effect. Appendix Figure E.4 presents the full outcome-by-outcome comparison.

Treatment effects are similar in magnitude across rural and urban schools. The CPS composite effect is 0.149 SD in rural schools (50 schools, $N = 2{,}775$) and 0.120 SD in urban schools (30 schools, $N = 1{,}613$); the interaction is not significant ($p_{\text{diff}} = 0.697$), though the uneven split limits power to detect a differential. Idea Generation effects are likewise comparable: 0.294 SD in rural versus 0.347 SD in urban ($p_{\text{diff}} = 0.773$). The similarity of point estimates across settings is reassuring from a scalability perspective. Appendix Figures E.5 and E.6 present the full outcome-by-outcome comparison by location and a summary of CPS composite heterogeneity across all four dimensions.

E. Idea Content: What Do Students Choose to Work On?

Beyond how well students innovate, this subsection examines *what* they choose to innovate on. In the Ideas Competition, each team selected an innovation sector (e.g., Agriculture, Environment, Education, Health) and identified a target beneficiary group. These choices reveal how students translate their training into real-world problem identification.

Table 7 reports sector choice distributions across four subgroup dimensions ($N = 1{,}283$ scored teams). Treatment marginally shifts sector choice (χ^2 $p = 0.080$): treated teams move away from Agriculture (-5.0 pp, $p < 0.05$) and toward Education ($+3.4$ pp, $p < 0.01$). The curriculum trains students to identify problems in their immediate surroundings; as residents of boarding schools, the school environment is their most salient daily context. The Find stage of the innovation cycle requires students to interview stakeholders and observe problems firsthand, and in a residential setting the people most accessible for these activities are teachers, peers, and school staff, making Education a natural domain for problem-finding.

Sector choice varies more strongly along demographic dimensions than by treatment status. Gender differences are highly significant (χ^2 $p = 0.001$): boys' teams disproportionately choose Environment (27% vs. 18%) and Transport, while girls' teams favor Social Welfare (15% vs. 10%) and Households. Rural–urban and tribal–social welfare differences are also highly significant ($p < 0.001$ in both cases), with rural and tribal welfare schools more likely to target Agriculture. Appendix Figures E.7–E.10 present the full sector and beneficiary distributions across all four dimensions. These patterns suggest that innovation content is shaped primarily by students' lived contexts and identities, with the curriculum providing a modest additional push toward novel problem domains.

VII. Discussion

This paper provides experimental evidence that a structured innovation curriculum can develop creative problem-solving among adolescents in under-resourced government schools. The curriculum raises a composite measure of creative problem-solving by 0.120 standard deviations, with effects on exploration behavior, idea generation, fluid intelligence, and problem solving confirmed across individual behavioral tasks, open-ended assessments, and team-based competitions evaluated by external judges. These gains are positive at every level of baseline academic ability; for the top quintile, the curriculum raises both creative problem-solving ($+0.258$ SD) and mathematics ($+0.218$ SD). In a year-end ideas competition scored blind by external evaluators, treatment teams score 0.223 SD higher on a validated innovation index, with the largest component effect on novelty. The gender heterogeneity is large: idea generation increases by 0.504 SD in girls' schools versus 0.095 SD in boys' schools ($p_{\text{diff}} = 0.010$), and innovation content varies systematically by demographics, consistent with evidence that inventor identity shapes the direction of innovation (Koning

TABLE 7—INNOVATION SECTOR CHOICE BY SUBGROUP (% OF TEAMS)

	Treatment		Gender		Location		School Type	
Sector	Control	Treat.	Boys	Girls	Urban	Rural	Social	Tribal
Environment	22.5	22.5	27.3	18.3	22.8	22.4	23.6	21.2
Agriculture	16.0	11.0	13.7	13.0	11.3	14.5	10.4	16.8
Crime & Safety	13.5	13.9	13.8	13.6	16.9	11.8	15.2	12.0
Social Welfare	13.0	12.1	9.7	14.9	13.8	11.7	10.9	14.4
Households	10.9	11.0	9.8	12.0	10.3	11.4	13.4	8.0
Education	4.6	8.0	6.3	6.6	7.3	6.0	6.4	6.5
Transport	5.3	7.3	7.3	5.6	7.7	5.6	5.2	7.9
Health	4.6	4.4	3.5	5.4	5.9	3.7	4.4	4.6
Other	9.6	9.6	8.5	10.5	4.0	12.9	10.4	8.6
N (teams)	586	697	600	683	478	805	699	584
χ^2 p-value	0.080		0.001		0.000		0.000	

Notes: Each cell reports the percentage of teams in that subgroup choosing the given sector. The unit of observation is a team (3–5 students) from the Ideas Competition ($N = 1,283$ teams). "Treatment" and "Control" refer to random assignment to the innovation curriculum. "Boys" and "Girls" refer to single-gender school type. "Urban" and "Rural" refer to school location. "Social" and "Tribal" refer to social welfare vs. tribal welfare residential schools. The χ^2 p-value tests the null hypothesis that sector choice is independent of each subgroup dimension (Pearson chi-squared test with $df = 8$).

et al., 2021; Einiö et al., 2023). These gains do not come at the cost of academic achievement: endline academic assessments and administrative board examination records show point estimates at or above zero.

These effects appear consistently across measures that differ in structure, incentives, and scoring. The exploration task is an individual incentivized game with no subjective evaluation. The creativity assessments are open-ended responses scored by a large language model trained on human ratings. The ideas competition is a team-based exercise evaluated blind by external judges. That treatment effects are positive across all three measurement families makes it unlikely that the gains reflect a single artifact: demand effects, scorer bias, or team dynamics. The consistency also addresses a practical obstacle to adopting higher-order skills programs in schools. Existing instruments for creativity and problem-solving were designed for small-sample laboratory settings, and policymakers have lacked scalable measures that could be embedded in routine evaluation. The multi-method strategy validated here, tablet-administered and scalable to government school settings, offers a path toward measuring what these programs produce.

The curriculum targets both cognitive skills and socio-emotional dimensions, but gains appear only on the cognitive side. Idea generation (+0.313 SD), problem solving (+0.142 SD), and fluid intelligence (+0.111 SD) all show significant effects. Emotional intelligence (+0.031 SD) and grit (+0.025 SD) do not, despite the curriculum explicitly training empathy through stakeholder interviews and persistence through iterative prototyping. Both are measured through behavioral performance tasks, not self-report: emotional intelligence through the Reading the Mind in the Eyes Test, which requires identifying emotions from photographs of eye regions, and grit through a Sudoku-based task in which students choose between easy and hard puzzles under real stakes. The emotional intelligence null is consistent with the distinction between practicing interpersonal skills in a team setting and recognizing emotions from static facial cues, a perceptual skill the curriculum does not directly train. The grit null is more informative: treated students *do* persist more effectively in the exploration task, searching more broadly and discovering higher-value outcomes over successive rounds, but they do not choose harder Sudoku puzzles at higher rates. The curriculum appears to build domain-specific persistence in innovation contexts rather than shifting a general preference for difficulty. Within the cognitive skills that do respond, effect sizes track dosage: idea generation, which receives the most direct practice through iterative brainstorming and prototyping, shows the largest gains; written expression, which the curriculum does not target, shows none. The treatment effect on fluid intelligence warrants caution.

Matrix reasoning is typically considered difficult to shift through educational interventions, and the effect here (0.111 SD) is modest. The result is consistent with sustained practice in structured problem decomposition transferring to abstract reasoning, but it does not establish that the curriculum improves fluid intelligence in the trait sense.

The curriculum replaced one 90-minute homework period per week with structured innovation activities over two academic years. On the Class X state board examinations, taken one year after the intervention ended and covering the near-universe of students in the original randomization, point estimates on all eight subjects are positive or zero. All three language subjects show marginally positive effects, consistent with the curriculum's emphasis on articulation: students brainstorm aloud, pitch ideas to peers, and justify design choices, activities that exercise expressive skills in a non-academic context. The no-displacement result is clear in the aggregate, but ceiling compression in the board examinations limits what can be learned about heterogeneity. The correlation between IRT ability and board exam mathematics scores is just 0.22, indicating that the exam does not discriminate well at the top of the distribution. The endline assessments show that the top quintile gains both creative problem-solving ($+0.258$ SD) and mathematics ($+0.218$ SD), but this complementarity cannot be confirmed on the board examination because the instrument compresses precisely the students for whom the effect is largest. Whether innovation training and academic performance are complements for the strongest students on high-stakes measures remains an open question that would require examinations with greater discriminating power at the upper tail.

The curriculum costs approximately forty dollars per student per year, covering materials kits, printing, mentoring, and end-of-year showcase events. This figure is comparable to other evidence-based interventions targeting non-routine skills: Alan and Mumcu (2024) develop curiosity at four dollars per child using printed teacher toolkits; Bando et al. (2019) improve mathematics through inquiry-based pedagogy at eighteen dollars per 0.10 SD gain; Chioda et al. (2026) develop business and negotiation skills through a residential program at $118 per participant. India's primary public investment in school-level innovation, the Atal Tinkering Lab program, provides grants of Rs 20 lakh (approximately $24,000) per school for equipment, infrastructure, and operating costs. For a school serving two hundred students, this implies a per-student cost roughly three times larger than the curriculum evaluated here. The government announced in 2025 that it would expand the program to 50,000 additional schools. No experimental evaluation of the Atal Tinkering Lab program has been conducted. The National Education Policy 2020 mandates the development of creativity and innovation across all schools; the results in this paper provide the first experimental evidence that such development is achievable within government schools, at a cost that is low relative to both existing public expenditure and the experimental benchmarks in the literature.

The evaluation was conducted in one of the most resource-constrained public school environments in India: government residential schools serving exclusively Scheduled Caste and Scheduled Tribe students, with no specialized equipment beyond a materials kit. That creative problem-solving gains emerge in this setting suggests that resource constraints alone are unlikely to prevent adoption elsewhere. Several features of the setting limit direct generalization. These are residential schools, which offer scheduling flexibility that day schools may not; they are single-gender, so the gender heterogeneity estimates capture classroom composition alongside gender; and the implementing organization provided beginning-of-year training and periodic monitoring, a support structure that may not replicate at scale without institutional investment. The similarity of treatment effects across rural and urban schools and across the ability distribution provides some reassurance, but whether the same curriculum produces comparable gains in co-educational day schools, or without external quality assurance, cannot be answered with these data. The question this paper does answer is whether the cognitive foundations of innovation are developable through structured training within the constraints of public schooling. The evidence establishes that they are. Whether these gains persist after the program ends, translate into downstream innovation or entrepreneurship, or generate labor market returns is an important question for future research. As routine cognitive skills become more rapidly automated (Acemoglu et al., 2026), the ability to develop non-routine skills at scale in populations that current education

systems do not target becomes a progressively more urgent policy problem. The results here suggest that the binding constraint is not talent or resources but structured exposure.

VIII. Conclusion

This paper asks whether a structured curriculum can develop the cognitive skills that underpin innovation within the constraints of public schooling. The evidence from a cluster randomized trial across 80 government residential schools in Telangana, India, indicates that it can. The curriculum produces gains in creative problem-solving, with effects appearing across the ability distribution, across measurement methods, and across school types. These gains do not come at the cost of academic achievement.

Several limitations constrain the interpretation of these findings. The study measures skill gains at the end of the intervention period; whether creative problem-solving gains persist after the curriculum ends is not measured. The intervention is a bundled treatment—curriculum content, trained student facilitators, structured group activities, and an end-of-year competition—and the design does not isolate which components drive effects. Whether developing these cognitive skills during adolescence leads to increased innovation, entrepreneurship, or productivity in adulthood remains an open question. Evidence from related interventions is encouraging: Chioda et al. (2026) find that youth business skills training in Uganda produces firm revenue gains nine years later; Ashraf et al. (2025b) show that teaching negotiation skills to adolescent girls in Zambia increases educational attainment a full decade later. Whether the specific skill gains documented here generate comparable long-run returns requires dedicated follow-up.

Cross-country differences in innovation output are typically attributed to institutional factors: intellectual property protections, research investment, market structure, and barriers to talent allocation (Acemoglu and Cao, 2012; Kerr and Nanda, 2014; Furman et al., 2002; Hsieh et al., 2019). These explanations describe the environment in which innovation occurs, not the process through which individuals develop the skills to innovate. The students in this study represent a population that is, by virtually any metric, among the furthest from the innovation frontier. Whether skill development in schools ultimately translates into innovation output depends on the persistence of these gains and the complementary institutions that convert cognitive skills into productive innovation. But the prior question—whether the cognitive foundations of innovation are malleable through formal education, even among the most disadvantaged student populations—is answered in this paper. They are.

References

Daron Acemoglu and Dan Cao. Innovation by entrants and incumbents. *Journal of Economic Theory*, 157: 255–294, 2012.

Daron Acemoglu, David Autor, and Simon Johnson. Building pro-worker artificial intelligence. Working Paper 34854, National Bureau of Economic Research, 2026.

Philippe Aghion, Ufuk Akcigit, Ari Hyytinen, and Otto Toivanen. The social origins of inventors. Technical report, National Bureau of Economic Research, 2017.

Sule Alan and Seda Ertac. Fostering patience in the classroom: Results from randomized educational intervention. *Journal of Political Economy*, 126(5):1865–1911, 2018.

Sule Alan and Ipek Mumcu. Nurturing childhood curiosity to enhance learning: Evidence from a randomized pedagogical intervention. *American Economic Review*, 114(4):1173–1210, 2024.

Sule Alan, Teodora Boneva, and Seda Ertac. Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *Quarterly Journal of Economics*, 134(3):1121–1162, 2019.

Michael L. Anderson. Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training projects. *Journal of the American Statistical Association*, 103(484):1481–1495, 2008.

Nava Ashraf, Abhijit Banerjee, and Vesall Nourani. Learning to teach by learning to learn. Working Paper, 2025a.

Nava Ashraf, Natalie Bau, Corinne Low, and Xiaoyue Shan. Ten years of relational power: The long-run effects of teaching negotiation skills to adolescent girls. Working Paper 34339, National Bureau of Economic Research, 2025b. Conditionally accepted, *AER: Insights*.

Aspiring Minds. National employability report: Engineers 2019. Technical report, Aspiring Minds, New Delhi, 2019.

Australian Curriculum, Assessment and Reporting Authority. General capabilities in the Australian curriculum. Technical report, ACARA, Sydney, 2013. URL https://www.australiancurriculum.edu.au/f-10-curriculum/general-capabilities/.

David H. Autor, Frank Levy, and Richard J. Murnane. The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, 118(4):1279–1333, 2003.

Rosangela Bando, Emma Näslund-Hadley, and Paul Gertler. Effect of inquiry and problem based pedagogy on learning: Evidence from 10 field experiments in four countries. NBER Working Paper No. 26280, 2019.

Simon Baron-Cohen, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb. 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, 2001.

Alexander Bell, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen. Who becomes an inventor in America? The importance of exposure to innovation. *Quarterly Journal of Economics*, 134(2): 647–713, 2019.

Eric Bettinger, Stein Ludvigsen, Mari Rege, Ingeborg F. Solli, and David Yeager. Increasing perseverance in math: Evidence from a field experiment in Norway. *Journal of Economic Behavior & Organization*, 146: 1–15, 2018.

Barbara Biasi and Song Ma. The education-innovation gap. *NBER Working Paper No. 29853*, 2022.

Robert L. Brennan. Generalizability theory. *Educational Measurement: Issues and Practice*, 11(4):27–34, 2001.

Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond. Generative AI at work. *Quarterly Journal of Economics*, 140(2), 2023.

Laura Chioda, Paul J. Gertler, David Contreras-Loya, and Dana Rose Carney. Making entrepreneurs: Long term returns to training youth in business skills. Working Paper 34637, National Bureau of Economic Research, 2026.

David M. Condon and William Revelle. The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, 43:52–64, 2014.

Creativity, Culture and Education. Creative partnerships: Changing young lives. Technical report, Arts Council England, London, 2012.

Delhi Government. Delhi's education revolution: Annual report 2019–20. Technical report, Directorate of Education, Government of NCT of Delhi, New Delhi, 2019.

David J. Deming. The growing importance of social skills in the labor market. *The Quarterly Journal of Economics*, 132(4):1593–1640, 2017.

David J. Deming. Four facts about human capital. *Journal of Economic Literature*, 60(3):839–866, 2022.

David J. Deming and Mikko Silliman. Skills and human capital in the labor market. Working Paper 32908, National Bureau of Economic Research, 2024.

Florian Ederer and Gustavo Manso. Is pay for performance detrimental to innovation? *Management Science*, 59(7):1496–1513, 2013.

Per-Anders Edin, Peter Fredriksson, Martin Nybom, and Björn Öckert. The right stuff? personality and labor market returns. *Economic Journal*, 132(642):1231–1261, 2022.

Elias Einiö, Josh Feng, and Xavier Jaravel. Social push and the direction of innovation. *Quarterly Journal of Economics*, 2023. Forthcoming.

Finnish National Board of Education. National core curriculum for basic education 2014. Technical report, Finnish National Board of Education, Helsinki, 2016. Implemented 2016; introduced transversal competences including creative thinking and multiliteracy.

Jeffrey L. Furman, Michael E. Porter, and Scott Stern. The determinants of national innovative capacity. *Research Policy*, 31(6):899–933, 2002.

Uri Gneezy and Jan Potters. An experiment on risk taking and evaluation periods. *Quarterly Journal of Economics*, 112(2):631–645, 1997.

James J. Heckman and Tim Kautz. Hard evidence on soft skills. *Labour Economics*, 19(4):451–464, 2012.

Santiago Hermo, Miika Päällysaho, David Seim, and Jesse M. Shapiro. Labor market returns and the evolution of cognitive skills: Theory and evidence. *Quarterly Journal of Economics*, 137(4):2309–2361, 2022.

Chang-Tai Hsieh, Erik Hurst, Charles I Jones, and Peter J Klenow. The allocation of talent and US economic growth. *Econometrica*, 87(5):1439–1474, 2019.

C. Kirabo Jackson, Shanette C. Porter, John Q. Easton, Alyssa Blanchard, and Sebastián Kiguel. School effects on socioemotional development, school-based arrests, and educational attainment. *American Economic Review: Insights*, 2(4):491–508, 2020.

Nam-Hwa Kang. STEAM education in Korea: Current policies and future directions. *Policy Futures in Education*, 17(2):226–244, 2019.

William R. Kerr and Ramana Nanda. Financing innovation. *Annual Review of Financial Economics*, 6(1): 361–384, 2014.

Jeffrey R. Kling, Jeffrey B. Liebman, and Lawrence F. Katz. Experimental analysis of neighborhood effects. *Econometrica*, 75(1):83–119, 2007.

Rembrand Koning, Sampsa Samila, and John-Paul Ferguson. 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, 2021.

David S. Lee. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(3):1071–1102, 2009.

Erik Lindqvist and Roine Vestman. The labor market returns to cognitive and noncognitive ability: Evidence from the swedish enlistment. *American Economic Journal: Applied Economics*, 3(1):101–128, 2011.

James G. March. Exploration and exploitation in organizational learning. *Organization Science*, 2(1):71–87, 1991.

Mercer Mettl. India's graduate skill index 2025. Technical report, Mercer Mettl, New Delhi, 2025.

Ministry of Education, Atal Innovation Mission, NITI Aayog. School innovation marathon 2024. https://schoolinnovationmarathon.org, 2024. Joint initiative of Ministry of Education, AIM-NITI Aayog, MIC-AICTE, and UNICEF YuWaah.

Ministry of Education, India. National education policy 2020. Technical report, Government of India, 2020.

Ministry of Education, Singapore. Applied learning programme. Technical report, Ministry of Education, Singapore, 2013. URL https://www.moe.gov.sg/education-in-sg/our-programmes/applied-learning.

Ministry of ICT and Innovation, Rwanda. National coding curriculum framework. Technical report, Government of Rwanda, 2019.

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

Karthik Muralidharan and Abhijeet Singh. Adapting for scale: Experimental evidence on technology-aided instruction in india. Working Paper 34205, National Bureau of Economic Research, 2025.

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

NITI Aayog. Atal innovation mission: Fostering innovation and entrepreneurship in schools. Technical report, Government of India, 2020. URL https://aim.gov.in/atal-tinkering-labs.php.

OECD. PISA 2022 results: Factsheets. Technical report, Organisation for Economic Co-operation and Development, Paris, 2023.

OECD. PISA 2022 results: Creative minds, creative schools. Technical report, OECD Publishing, Paris, 2024.

OECD and Eurostat. *Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation*. OECD Publishing, Paris, 4th edition, 2018.

Abhijeet Singh. Learning more with every year: School year productivity and international learning divergence. *Journal of the European Economic Association*, 18(4):1770–1813, 2020.

Abhijeet Singh. Myths of official measurement: Auditing and improving administrative data in developing countries. *Review of Economics and Statistics*, 106(1):42–56, 2024.

Jivika Soni and Kunal Sharma. The shark tank india effect: Surge in entrepreneurship and startup pitches among school and college students. *The Economic Times*, February 2022. URL https://economictimes.indiatimes.com/tech/startups/the-shark-tank-india-effect-62-surge-in-entrepreneurship-startup-pitches-among-school-college-articleshow/89714378.cms.

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

Martin R. West, Matthew A. Kraft, Amy S. Finn, Rebecca E. Martin, Angela L. Duckworth, Christopher F. O. Gabrieli, and John D. E. Gabrieli. Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis*, 38(1):148–170, 2016.

Wheebox, CII, and AICTE. India skills report 2023. Technical report, Wheebox, 2023. Annual employability survey of Indian graduates.

World Bank. World development report 2018: Learning to realize education's promise. *World Bank Publications*, 2018.

World Economic Forum. The future of jobs report 2025. Technical report, World Economic Forum, Geneva, 2025.

David S. Yeager, Paul Hanselman, Gregory M. Walton, Jared S. Murray, Robert Crosnoe, Chandra Muller, et al. A national experiment reveals where a growth mindset improves achievement. *Nature*, 573(7774): 364–369, 2019.

Appendices

A. The Intervention: Curriculum Detail

A. Session-Level Audit

TABLE A.1—THINK & MAKE PROGRAM — ALL SESSIONS

Year	Block	Unit Theme	Session	FLIMP Stage	Activity Description	Primary Skills
Year 1, Block 1 — Planet (Environmental Problems)						
1	1	Planet	1	—	Introduction to innovation; distinguishing problems from solutions; community mapping	Observation, emotional intelligence
1	1	Planet	2	Find	Guided story card: identify environmental problem; interview stakeholders; cause–effect mapping	Emotional intelligence
1	1	Planet	3	Learn	Study real-world case studies of environmental solutions; analyze trade-offs	Fluid intelligence, reasoning
1	1	Planet	4	Ideate	Brainstorm solutions using structured templates; evaluate against feasibility criteria	Divergent thinking, creativity
1	1	Planet	5	Make	Build prototype from low-cost materials; test and iterate	Problem solving, grit
1	1	Planet	6	Make	Continue prototyping; incorporate peer feedback; refine design	Problem solving, collaboration
1	1	Planet	7	Present	Prepare pitch; present to class panel; defend design choices	Communication, confidence
1	1	Planet	8	Present	Reflection and documentation; showcase preparation	Communication, reflection
Year 1, Block 2 — People (Social Problems)						
1	2	People	1	—	Revisit FLIMP process; introduction to social problems; empathy exercises	Emotional intelligence, observation
1	2	People	2	Find	Guided story card: identify social/community problem; stakeholder interviews	Emotional intelligence
1	2	People	3	Learn	Analyze existing social interventions; evaluate evidence and trade-offs	Fluid intelligence, reasoning
1	2	People	4	Ideate	Generate solution concepts; rank by impact and feasibility	Divergent thinking, creativity
1	2	People	5	Make	Build prototype; test with intended users	Problem solving, grit
1	2	People	6	Make	Iterate on prototype based on user feedback	Problem solving, collaboration
1	2	People	7	Present	Team pitch to evaluator panel; Q&A defense	Communication, confidence
1	2	People	8	Present	Reflection; cross-team feedback; showcase	Communication, reflection
Year 2, Block 1 — Design Evolutions (Scaffolding Gradient)						
2	3	Design Evolutions	1	Find	Guided story card (e.g., chalk waste): observe problem, interview stakeholders, map causes	Emotional intelligence, observation

Continued on next page

Table A.1 continued from previous page

Year	Block	Unit Theme	Session	FLIMP Stage	Activity Description	Primary Skills
2	3	Design Evolutions	2	Learn–Ideate	Study case solutions; brainstorm and evaluate alternatives	Fluid intelligence, creativity
2	3	Design Evolutions	3	Make–Present	Build and present guided-card solution; class critique	Problem solving, communication
2	3	Design Evolutions	4	Find–Ideate	Semi-open quick challenge: teacher provides problem domain; students scope and ideate independently	Fluid intelligence, divergent thinking
2	3	Design Evolutions	5	Make–Present	Build and present quick-challenge solution	Problem solving, communication
2	3	Design Evolutions	6	Find–Make	Independent problem discovery: students identify, scope, and prototype their own problem	All FLIMP skills
2	3	Design Evolutions	7	Present	Final presentation and reflection; peer evaluation	Communication, reflection
Year 2, Block 2 — Smarter Everyday (Sensors & Technology)						
2	4	Smarter Everyday	1	—	Introduction to sensors: human senses → electronic sensors; hands-on circuit building with Input Board, 4 outputs (LED, motor, propeller, buzzer), 6 sensors (sound, heat, distance, touch, light, water)	Reasoning, observation
2	4	Smarter Everyday	2	Learn	Real-world sensor applications via case study cards: touch sensor (ATMs), motion sensor (security), distance sensor (smart doors), pressure sensor (tires), LPG sensor (gas leaks)	Fluid intelligence, reasoning
2	4	Smarter Everyday	3	Find–Ideate	Guided story card: street lights always on during daytime; interview stakeholders; ideate sensor-based solutions	Emotional intelligence, creativity
2	4	Smarter Everyday	4	Make–Present	Build sensor-based prototype for story card problem; present solution and receive critique	Problem solving, communication
2	4	Smarter Everyday	5	Find	Independent problem discovery: identify real-world problem amenable to sensor-based solution; stakeholder interviews	Emotional intelligence, observation
2	4	Smarter Everyday	6	Ideate–Make	Design and build independent sensor-integrated prototype; test and iterate	Creativity, problem solving
2	4	Smarter Everyday	7	Present	Final team presentation; defend design; reflection and showcase	Communication, confidence

Notes: Each row corresponds to one 90-minute session of the Think & Make program. The program spans two academic years (four blocks), with each block following the five-stage FLIMP cycle: Find, Learn, Ideate, Make, Present. Sessions marked "—" are introductory or orientation sessions that precede the FLIMP cycle. Year 1 delivers two guided cycles on environmental and social problems. Year 2 Block 1 transitions from guided story cards through semi-open quick challenges to fully independent problem discovery within a single block. Year 2 Block 2 introduces electronic sensors and repeats the guided-to-independent progression with technology-integrated prototyping. Primary skills listed reflect the dominant cognitive demands of each session; most sessions engage multiple skills simultaneously. Skill labels follow the curriculum and outcome mapping in Figure 1. Source: Student workbooks, implementation tracker, and Product Note session–skill mapping (see Section II).

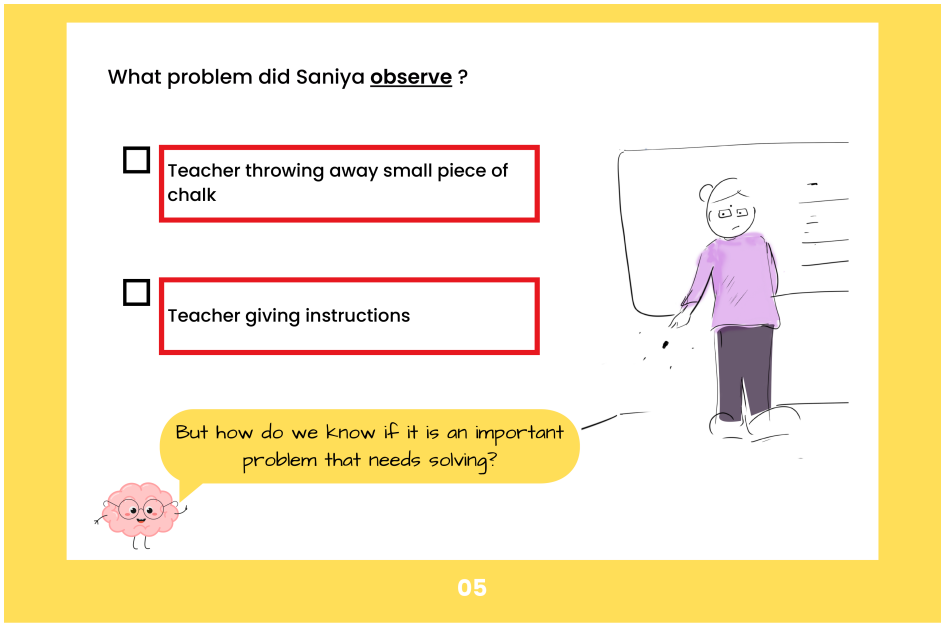
B. Skills Targeted by the Curriculum

TABLE A.2—SKILLS TARGETED BY THE CURRICULUM

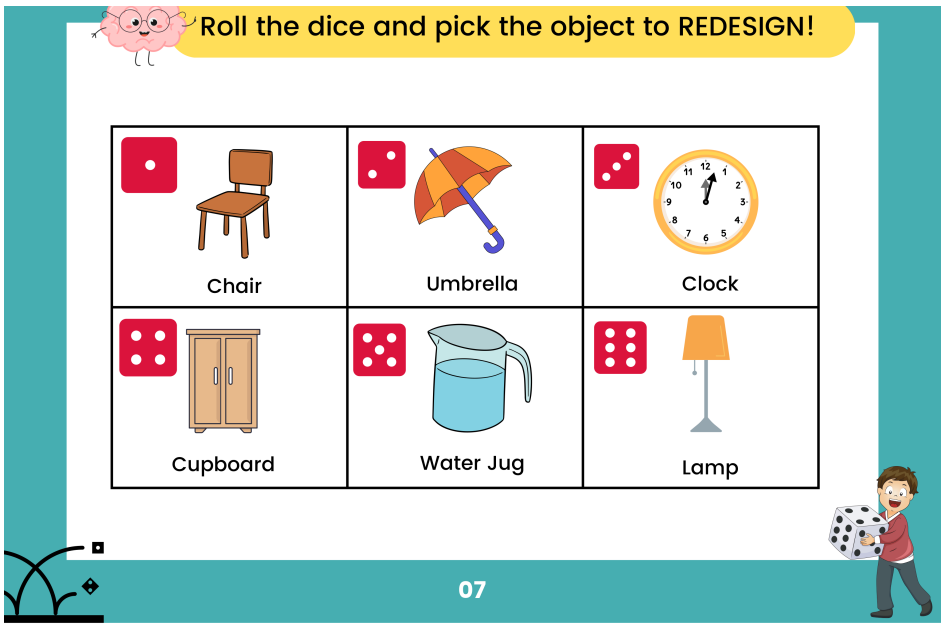
Skill	Definition	FLIMP Stage(s)	Sessions (per 16)	Curriculum Examples	Cognitive Demand
Fluid Intelligence	Systematic evaluation of evidence and arguments to form reasoned judgments	Learn, Ideate, Make	10	Cause–effect diagrams for chalk waste (Y2B1); evaluating trade-offs across environmental solutions (Y1B1); multi-criteria ranking of sensor designs (Y2B2)	Separating correlation from causation; weighing competing evidence; forming judgments under uncertainty
Problem Solving	Identifying, analyzing, and resolving novel problems through iterative strategies	Learn, Ideate, Make	10	Scoping ill-defined school-based problems (Y1); prototyping under material constraints (all blocks); debugging failed sensor circuits (Y2B2)	Decomposing novel problems; generating and testing hypotheses through prototypes; adapting strategies after failure
Creativity	Generating original, diverse, and contextually appropriate ideas	Ideate, Make	8	Constrained brainstorming with templates (Y1); dice-based rapid design challenges (Y2B1); adapting sensor solutions across domains (Y2B2)	Divergent production under constraints — fluency, flexibility, and originality
Emotional Intelligence	Understanding others' experiences by adopting their viewpoint	Find, Ideate	8	Stakeholder interviews (all blocks); story cards — chalk waste, street lights (Y2); designing for specific user groups	Perspective-taking, empathic concern, and theory of mind
Grit	Sustained effort and adaptive response to setbacks	Make	—	Multi-session prototype builds with iterative failure (Y1); redesigning after peer critique (all blocks); debugging sensor circuits (Y2B2)	Persistence through repeated failure; willingness to restart; sustained engagement across sessions
Collaboration	Coordinating effectively within a team toward a shared goal	All stages	10	Permanent teams of 4; division of labor in prototyping; collective ideation and consensus-building; joint presentations	Negotiation, role coordination, shared mental models
Communication	Conveying ideas clearly through oral and written expression	Present, discus-sions	8	Structured pitch preparation (all blocks); Q&A defense before panels; written design documentation in workbooks	Organizing ideas for diverse audiences; responding to critical questions in real time
Presentation	Organizing and delivering a persuasive argument to an audience	Present	10	Formal team presentations at end of each FLIMP cycle; peer feedback sessions; showcase events (Y1B2, Y2B2)	Structured argumentation, audience awareness, composure under questioning
Creative Confidence	Belief in one's ability to generate and implement creative ideas	Learn	4	Age-appropriate innovator role models in case studies (Y1); scaffolded early successes before independent work; progressive autonomy across blocks	Self-efficacy through mastery experiences — early successes support later risk-taking

Notes: Columns report curriculum-to-skill mapping only; measurement instruments and psychometric properties are documented in Appendix C. Sessions of Exposure counts the number of sessions (out of 16 per two-block cycle) in which the skill is a primary target, from Product Note Table 3. Grit is exercised primarily through the Make stage but was not assigned a session count in the implementation documentation. Year–Block abbreviations: Y1B1 = Year 1 Block 1, Y2B1 = Year 2 Block 1, Y2B2 = Year 2 Block 2.

C. *Illustrative Curriculum Materials*



Panel A: Guided — Story-card comprehension check (Year 2, Block 1). After reading a narrative about Saniya observing chalk waste in her school, students answer a structured question with predefined options. The workbook then prompts them to evaluate whether the identified problem is worth solving.



Panel B: Semi-Open — Dice-based random design challenge (Year 2, Block 1). Students roll a die to receive a randomly assigned everyday object (chair, umbrella, clock, cupboard, water jug, or lamp), then redesign it for a randomly assigned user persona. The object is constrained; the redesign is open-ended.

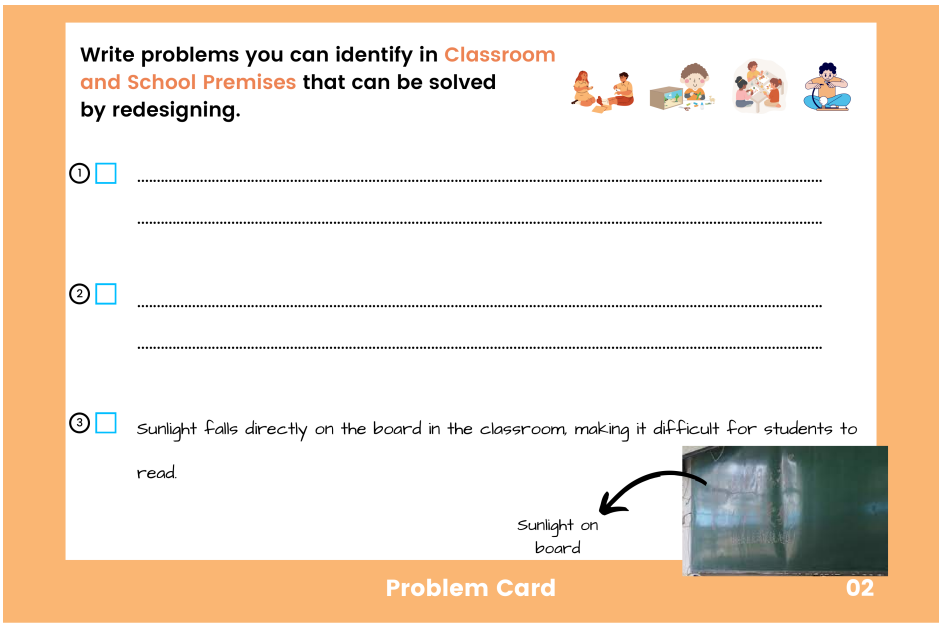


FIGURE A.1. ILLUSTRATIVE WORKBOOK PAGES ACROSS THE SCAFFOLDING GRADIENT

Panel C: Independent — Problem-discovery card (Year 2, Block 1). Students identify problems in their classroom and school premises that could be solved by redesigning an object. Only one example is provided (item 3); items 1 and 2 are entirely student-generated. Subsequent cards repeat this format for home, everyday objects, and public places.

Notes: Three panels illustrate the progressive release of autonomy described in Section II. Panel A shows a guided comprehension check where students identify a problem from a structured narrative. Panel B shows a semi-open challenge where the design object is randomly assigned but the solution is open-ended. Panel C shows a fully independent problem-discovery page where students identify, scope, and describe their own problems with minimal scaffolding. Source: Student workbooks, reproduced with permission from Inqui-Lab Foundation.

D. Delivery Model and Counterfactual

Facilitators and Materials

Sessions were facilitated by two student leaders per classroom, nominated by teachers from among the participating students. Leaders received approximately 20 hours of training from Inqui-Lab Foundation staff before each academic year, covering session logistics, material handling, timekeeping, and team coordination. Crucially, leaders were instructed to answer only clarifying questions—they did not lecture, explain concepts, or evaluate student work. This design preserved students' ownership of the learning process and ensured that any treatment effects could not be attributed to instructional quality.

Each classroom received a standardized innovation kit ("Curiosity Box") containing circuits, sensors, recyclable materials, and beginner-friendly prototyping tools aligned to the curriculum's technical demands. All sessions were delivered through structured student workbooks containing step-by-step prompts, reflection activities, and peer feedback mechanisms. The workbooks enabled teams to collaborate and make progress semi-autonomously, reducing dependence on facilitator quality. Leaders participated in debriefing sessions after each curriculum block to discuss implementation challenges and standardize delivery across schools.

Implementation Fidelity

Implementation fidelity was monitored through weekly observations by a field staff team of seven trained members. Observers randomly sampled schools and sessions and used standardized rubrics to assess completion of tasks, facilitation quality, and student engagement. In Year 1, 35.2% of sessions were observed; in Year 2, 37.5% were observed. Across all 40 treatment schools, 98.1% completed every required session, and 88.0% of observed sessions implemented all core components.

Counterfactual Conditions

Control schools participated in the same annual Innovation Challenge as treatment schools. At the beginning of each academic year, both treatment and control schools received a formal letter notifying them of an upcoming inter-school competition (the "Shark Tank Challenge"), emphasizing that teams would be evaluated based on how beneficial their innovative idea was for its intended users. The letter provided basic competition logistics but offered no guidance on idea development or preparation. Teachers in control schools conducted regular check-ins with student teams to ensure they were developing products by the same deadlines as treatment teams. Both groups submitted standardized documentation of their products (photographs and descriptions) via an online survey.

The key treatment-control contrast is therefore structured training versus unstructured opportunity: treatment teams received approximately 45 contact hours of curriculum-based instruction following the FLIMP cycle, while control teams received the same competitive incentive and deadlines but no pedagogical support, materials, or facilitated sessions. Sessions occurred during after-school hours in these residential schools, displacing time that students would otherwise spend on homework or independent study. Endline time-use data confirm that control students allocated similar shares of their non-school time to homework (19%), sports (19%), innovation-related activities (17%), and other activities (20%), indicating a diffuse counterfactual rather than a single dominant substitute activity.

B. Design and Data

A. Attrition Analysis

TABLE B.1—ATTRITION ANALYSIS

	Treatment	Control	Difference	(SE)	N
Panel A: Survey Response Rates					
Baseline Survey	70.9	71.9	−0.9	(2.5)	5,642
Midline Survey	81.8	82.0	−0.1	(1.5)	5,642
Endline Survey	79.8	75.7	4.1	(2.6)	5,642
Class X Board Exams	99.9	99.8	0.1	(0.1)	5,642
Panel B: Predictors of Endline Attrition					
	Covariate Level		Treatment × Covariate		
	β	(SE)	β	(SE)	
Treatment	0.0238	(0.1844)			
Age	−0.0029	(0.0081)	−0.0031	(0.0129)	
Household Assets Index	0.0057	(0.0064)	−0.0087	(0.0078)	
Computer Literacy	0.0347	(0.0299)	−0.0787**	(0.0359)	
Baseline Mathematics	0.0106	(0.0110)	−0.0139	(0.0158)	
Baseline Science	0.0044	(0.0141)	0.0168	(0.0191)	
Baseline Fluid Intelligence	−0.0086**	(0.0034)	0.0031	(0.0046)	
Baseline Emotional Intelligence	−0.0072**	(0.0029)	−0.0012	(0.0036)	
Joint F-test (T×X)		$F = 1.53$, $p = 0.149$			
Mean attrition rate		22.2%			
Class X Board Exam Scores by Endline Attendance					
	β(Attrited)	(SE)	p-value		N
Class X Mathematics	0.036	(0.065)	0.580		5,632
Class X Science	−0.026	(0.045)	0.572		5,632
Panel C: Lee (2009) Bounds on Main Outcomes					
	Naïve ITT	Lower Bound	Upper Bound		N
Exploration Index	0.057	−0.010	0.168		4,388
Idea Generation	0.310	0.159	0.366		4,388
Fluid Intelligence	0.108	0.020	0.212		4,388
Problem Solving	0.102	−0.005	0.190		4,388
Mathematics	−0.044	−0.152	0.053		4,388
Science	0.008	−0.110	0.101		4,388

Notes: Panel A reports survey response rates by treatment status. The randomized sample comprises 5,642 students in 80 schools (40 treatment, 40 control). Baseline survey was administered before treatment assignment; midline after Year 1; endline after Year 2. Class X board examination scores are administrative records with near-complete coverage. Differences are estimated from a regression of presence on treatment with strata fixed effects; standard errors clustered at school level. Panel B reports coefficients from a regression of endline attrition on treatment, baseline covariates, treatment × covariate interactions, and strata fixed effects. The joint F-test tests the null that all treatment × covariate interactions are jointly zero, i.e., that attrition is not differentially selective across treatment arms. The Class X rows regress control-group-standardized board exam external marks on an endline-attrition indicator with strata fixed effects and clustered standard errors, testing whether endline attenders and non-attenders perform similarly on the nearly attrition-free benchmark ($N \approx 5{,}632$). Panel C reports Lee (2009) trimming bounds. Treatment has 79.8% endline presence vs. 75.7% for control; the top/bottom 5.1% of treatment outcomes are trimmed to equalize group sizes. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

B. Tablet-Based Assessment Protocol

All assessments were administered on individual tablets using the Qualtrics Offline App, which enabled survey delivery in schools without reliable internet connectivity. Enumerators were organized into teams of five, each carrying pre-charged tablets in protective cases. At each school, enumerators registered students by name and unique ID, distributed tablets, and supervised the session in a designated classroom. Each student received a tablet preloaded with the complete assessment battery in both English and Telugu. The order of assessment modules was randomized across students to mitigate survey fatigue, ordering effects, and copying—students seated next to one another encountered different modules at any given time. Each question appeared on a separate screen to further limit cheating.

Enumerators received three days of training covering device logistics, troubleshooting, and protocols for handling student questions (they were instructed to answer only technical queries about the interface, not content). At the end of each school visit, response data were synced to the Qualtrics server over mobile hotspot and verified for completeness by the field coordinator. All instruments were piloted in two out-of-sample schools before being administered to the study sample; pilot feedback led to minor adjustments in item ordering and Telugu translations.

C. Control Variables

The main specification follows Kling et al. (2007) and includes the following baseline controls:

1. **Lagged dependent variable:** The outcome-specific baseline measure (e.g., baseline ICAR matrix reasoning theta for the endline fluid intelligence regression; baseline math theta for the endline mathematics regression). For outcomes measured only at endline (creativity, grit, problem solving), the closest baseline proxy is used (e.g., baseline exploration index for creativity outcomes).
2. **Missing-data indicator:** A binary flag equal to one for students who did not complete the baseline assessment ($N = 1{,}613$, or 28.6% of the randomization roster). For these students, the lagged dependent variable is imputed at the control-group mean.
3. **Individual characteristics:** Age at baseline, a PCA-based socioeconomic status index (first principal component of eight binary household asset indicators: air conditioner, car, refrigerator, laptop, scooter, washing machine, father owns mobile phone, mother owns mobile phone; PC1 explains 29.8% of variation), and computer literacy (binary).
4. **School infrastructure (UDISE):** Nine variables drawn from the Unified District Information System for Education administrative records: building ownership (owned vs. rented), boundary wall condition (pucca, partial, or none), number of functional toilets, and binary indicators for library, playground, medical facility, English medium of instruction, and computer availability.
5. **Strata fixed effects:** Eight indicators for the randomization strata (management type $\times$ student composition $\times$ location).

C. Measurement

A. Innovation Index: Scale Construction and Psychometrics

TABLE C.1—INNOVATION INDEX: SCALE CONSTRUCTION AND PSYCHOMETRIC PROPERTIES

	Dimension	Mean	SD	$r_{i,t}$	α if deleted
Panel A: Scale Validation on Adult Innovators ($N = 40$)					
Offers a new solution	Novelty	4.45	0.78	0.386	0.396
Improves upon existing solution		3.98	1.33	0.278	0.406
Combines existing solutions		4.12	1.09	0.250	0.438
Transplants solution across contexts		3.77	1.33	0.321	0.385
Solves problem affecting many	Value-Add	4.47	0.64	0.380	0.414
Improves lives of users		4.12	1.04	−0.076	0.576
Easily accessed by those in need		4.55	0.75	0.245	0.447
Cronbach's α (7 items)			0.485		
Novelty subscale (4 items)			0.546		
Value-Add subscale (3 items)			0.393		
Panel B: Student Ideas Competition ($N = 1{,}189$ graded teams)					
Offers a new solution	Novelty	3.64	1.16	0.841	0.942
Improves upon existing solution		3.92	0.97	0.880	0.937
Combines existing solutions		3.74	0.99	0.876	0.938
Transplants solution across contexts		3.68	0.97	0.878	0.938
Solves problem affecting many	Value-Add	3.93	0.92	0.763	0.947
Improves lives of users		3.52	1.01	0.770	0.947
Easily accessed by those in need		4.13	0.85	0.826	0.943
Cronbach's α (7 items)			0.950		
Novelty subscale (4 items)			0.938		
Value-Add subscale (3 items)			0.842		
First eigenvalue			5.42 (77% of variance)		

Notes: The Innovation-S scale comprises 7 items measuring two dimensions of innovation quality: Novelty (4 items) and Value-Add (3 items). Panel A reports psychometric properties from a validation sample of adult innovators (entrepreneurs, product managers, and researchers) who rated their own innovations. Panel B reports properties from the student ideas competition, where two blind evaluators independently scored each team's submission; reported statistics use evaluator-mean scores. $r_{i,t}$ denotes the corrected item-total correlation (item excluded from total). α if deleted is Cronbach's alpha when the item is removed from the scale.

TABLE C.2—INNOVATION INDEX: INTER-RATER RELIABILITY AND INDEX WEIGHTING

Panel A: Inter-Rater Reliability

	Dimension	Pearson r	N
Novelty 1	Novelty	0.239	1,151
Novelty 3	Novelty	0.214	1,151
Novelty 4	Novelty	0.232	1,151
Novelty 5	Novelty	0.202	1,151
Value-Add 3	Value-Add	0.238	1,151
Value-Add 4	Value-Add	0.165	1,151
Value-Add 5	Value-Add	0.205	1,151
Feasibility		0.158	1,151
Funding		0.348	1,151

Panel B: Anderson (2008) Inverse-Covariance Weights

	Dimension	ICW Weight
Offers a new solution	Novelty	−0.203
Improves upon existing solution	Novelty	0.079
Combines existing solutions	Novelty	0.005
Transplants solution across contexts	Novelty	0.063
Solves problem affecting many	Value-Add	0.168
Improves lives of users	Value-Add	0.358
Easily accessed by those in need	Value-Add	0.530

Notes: Panel A reports Pearson correlations between Evaluator 1 and Evaluator 2 raw item scores for the 1,151 teams scored by both evaluators. Feasibility and Willingness to Fund are single-item measures reported alongside the 7-item Innovation-S scale. All student-level indices use within-evaluator z-scoring to remove rater severity effects, followed by control-group standardization ($N_C = 563$ graded control teams). The primary Innovation Index weights Novelty and Value-Add equally (50/50 dimension weighting). Panel B reports Anderson (2008) inverse-covariance weights, estimated from the graded control group's covariance matrix, as a robustness alternative that upweights items contributing unique information.

B. Exploration Task: Game Design and Payoff Structure

TABLE C.3—EXPLORATION TASK: GAME DESIGN AND PAYOFF STRUCTURE

Panel A: Stand Payoff Parameters

	Fresh Juice	Pani Puri	Ice Cream
Maximum profit (Rs.)	120	180	250
Penalty per unit deviation (Rs.)	10	20	30
Zero-profit combinations (%)	25.7	54.0	63.3
Mean profit (Rs.)	31.1	26.6	24.0
SD of profit (Rs.)	27.0	36.6	43.7
Risk–return profile	Low	Medium	High

Panel B: Choice Dimensions (per round)

	Options	Range	Optimal (FJ / PP / IC)
Stand type	3	Fresh Juice, Pani Puri, Ice Cream	—
Location	3	Market, Homes, School	1 / 2 / 3
Working hours	10	6–15 hours	8 / 7 / 9
Menu items	10	3–12 items	4 / 6 / 5
Selling price	10	Rs. 11–Rs. 20	5 / 3 / 4

Panel C: Task Structure

Number of rounds	12
Total possible combinations	9,000
Overall zero-profit rate	47.7%
Payoff structure	Deterministic (no noise)
Feedback	Immediate profit after each round
Default option profit	Rs. 30 (Fresh Juice, Market, 9 hrs, 5 items, Rs. 19)

Notes: The Exploration Task is a lab-in-the-field game adapted from Ederer and Manso (2013). Students inherit a street food business and make five decisions per round to maximize profit. Profit is $\pi_s(\mathbf{c}) = \max\{0,\, V_s - \lambda_s \sum_d |c_d - c^*_{s,d}|\}$, where V_s is the maximum profit, λ_s the penalty rate, and $c^*_{s,d}$ the optimal input. Students do not know these parameters and must discover profitable combinations through experimentation. The default option yields Rs. 30/round; systematic exploration is required to exceed this. Panel A statistics are computed over the full payoff landscape (9,000 combinations).

C. Exploration Task: Round-by-Round Treatment Effects

TABLE C.4—EXPLORATION TASK: ROUND-BY-ROUND TREATMENT EFFECTS

	Best Profit Found		Strategies Tried		Proximity to Optimal	
Round	$\hat{\beta}$	(SE)	$\hat{\beta}$	(SE)	$\hat{\beta}$	(SE)
1	-0.005	(0.037)	0.000	(0.000)	-0.008	(0.035)
2	0.024	(0.039)	0.053	(0.036)	-0.038	(0.036)
3	0.036	(0.041)	0.017	(0.038)	-0.045	(0.036)
4	0.035	(0.041)	0.039	(0.038)	-0.056	(0.037)
5	0.065	(0.041)	0.054	(0.039)	-0.079**	(0.038)
6	0.042	(0.040)	0.068*	(0.040)	-0.061*	(0.036)
7	0.048	(0.040)	0.083**	(0.041)	-0.060*	(0.036)
8	0.066	(0.041)	0.077*	(0.043)	-0.082**	(0.038)
9	0.075*	(0.042)	0.083*	(0.044)	-0.089**	(0.037)
10	0.087**	(0.043)	0.078*	(0.044)	-0.097**	(0.039)
11	0.072*	(0.043)	0.078*	(0.044)	-0.089**	(0.039)
12	0.066	(0.043)	0.077*	(0.043)	-0.086**	(0.040)

Notes: Each cell reports the OLS treatment effect on the cumulative outcome through round r, standardized to the control-group mean and SD. Best Profit Found is the maximum single-round profit discovered through round r. Strategies Tried is the number of unique (stand, location) combinations explored through round r. Proximity to Optimal is the closest Manhattan distance to any stand's optimal input vector through round r (lower values indicate closer proximity). All regressions include strata fixed effects (tribal $\times$ girls-only $\times$ rural), baseline individual controls (age, household assets index, computer literacy), 9 school infrastructure controls, and the baseline analog of the outcome. Standard errors clustered at the school level (80 clusters) in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

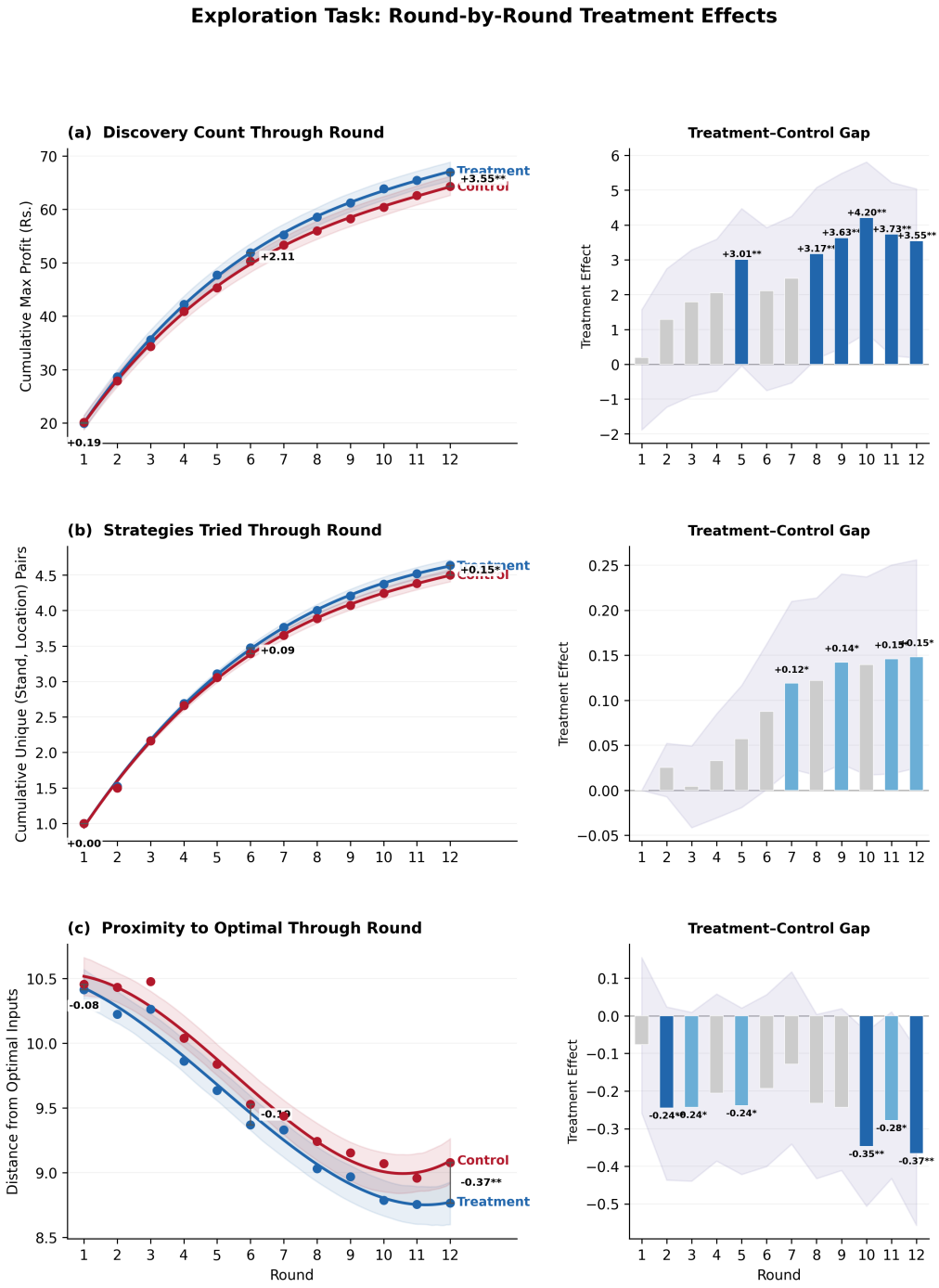


FIGURE C.1. EXPLORATION TASK — ROUND-BY-ROUND TREATMENT EFFECTS

Notes: Three outcomes shown in rows: (a) Discovery Count, (b) Strategies Tried, (c) Proximity to Optimal. Left panels: treatment and control means with cubic polynomial fits and 95% bootstrap CIs. Right panels: round-specific treatment–control gaps from OLS with strata FE, individual controls, infrastructure controls, and baseline analogs; SEs clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. $N \approx 4{,}388$.

D. Exploration Task: Payoff Landscape

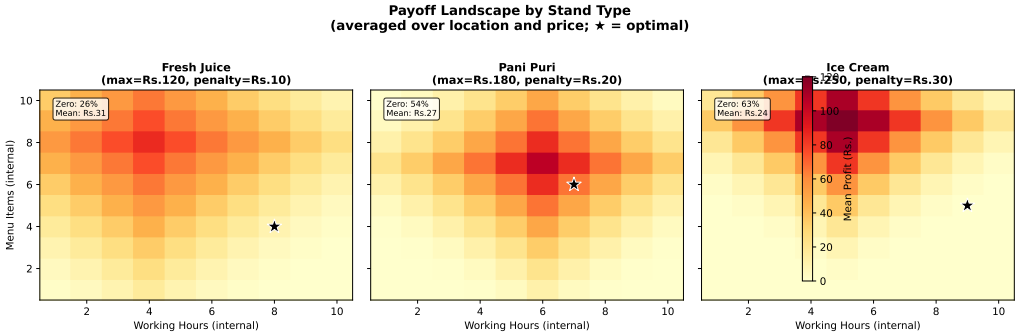


FIGURE C.2. PAYOFF LANDSCAPE BY STAND TYPE

Notes: Each panel shows the mean profit (in Rs.) across the working hours $\times$ menu items grid for one of the three stand types, averaged over all location and price combinations. The star marks the optimal input combination for each stand. Profit is computed as $\pi_s(\mathbf{c}) = \max\{0,\, V_s - \lambda_s \sum_d |c_d - c^*_{s,d}|\}$, where V_s is the maximum profit, λ_s the penalty rate, and $c^*_{s,d}$ the optimal value for input dimension d (Ederer and Manso, 2013). Fresh Juice has the lowest maximum profit (Rs. 120) but the gentlest penalty (Rs. 10 per unit deviation), while Ice Cream has the highest maximum (Rs. 250) but the steepest penalty (Rs. 30 per unit deviation), creating a risk–return tradeoff across stands. The total strategy space comprises 9,000 unique combinations across all three stands.

E. Fluid Intelligence: Sample Item

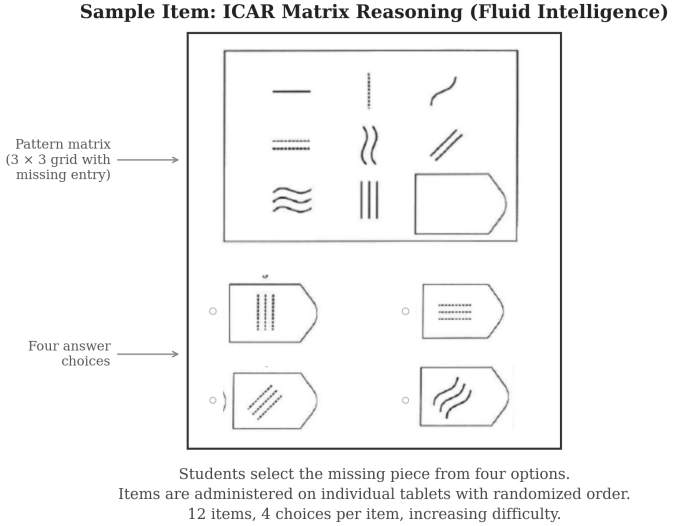


FIGURE C.3. SAMPLE ITEM: ICAR MATRIX REASONING (FLUID INTELLIGENCE)

Notes: Each item presents a 3×3 matrix of geometric patterns with one cell missing. The student selects the answer that completes the pattern from four options. The endline assessment comprises 12 items of increasing difficulty, administered on individual tablets with randomized item order. Items are scored as binary correct/incorrect. Ability scores are estimated using a two-parameter IRT model (test reliability = 0.981). Items are adapted from the International Cognitive Ability Resource (Condon and Revelle, 2014), a public-domain matrix reasoning battery validated against the WAIS-IV ($r = .81$), with answer options reduced from six to four for tablet readability.

F. Emotional Intelligence: Sample Item

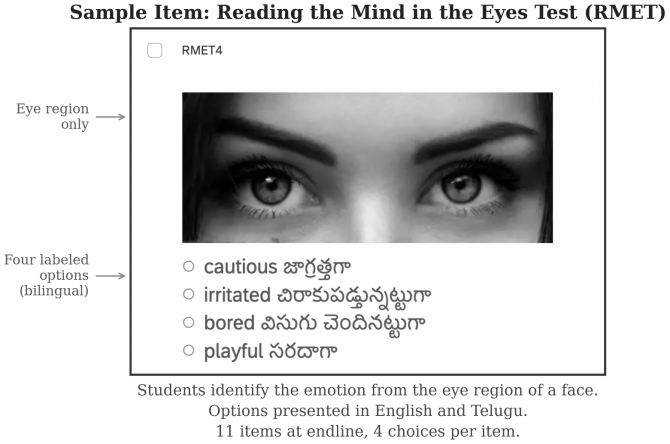


FIGURE C.4. SAMPLE ITEM: READING THE MIND IN THE EYES TEST (EMOTIONAL INTELLIGENCE)

Notes: Each item presents a photograph of the eye region of a face. The student selects the emotion that best describes what the person is feeling or thinking from four labeled options, presented in both English and Telugu. The endline assessment comprises 11 items administered on individual tablets with randomized item order. Items are scored as binary correct/incorrect. Ability scores are estimated using a two-parameter IRT model (test reliability = 0.933). The original RMET (Baron-Cohen et al., 2001) was adapted for the Indian context: stimuli were replaced with culturally appropriate photographs, and the adapted instrument was piloted with 150 out-of-sample students before deployment in the study.

G. Board Exam Ceiling Effects

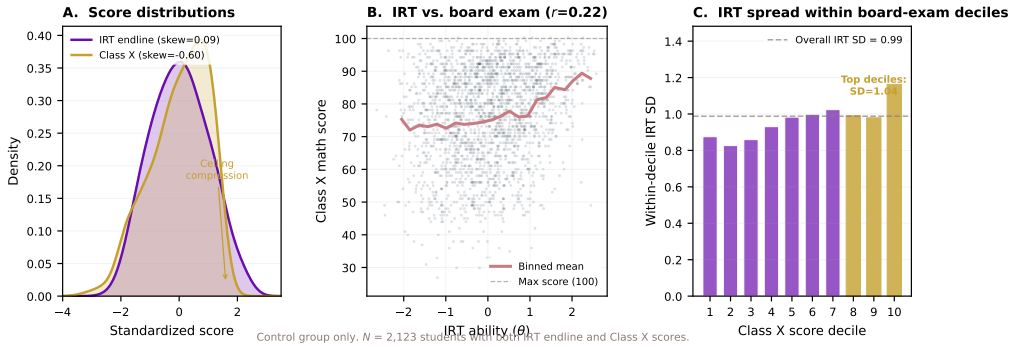


FIGURE C.5. CEILING COMPRESSION IN CLASS X BOARD EXAMS: MATHEMATICS

Notes: Three-panel figure for control-group students only ($N = 2{,}123$). Left panel: density comparison of IRT theta (symmetric) versus Class X z-score (left-skewed, ceiling-compressed). Center panel: scatter of IRT theta versus Class X raw score, showing fan-shape compression at the top of the board exam scale. Right panel: conditional standard deviation of IRT ability by Class X score decile, rising sharply at the top. The Pearson correlation between the IRT theta and Class X raw score is $r = 0.22$.

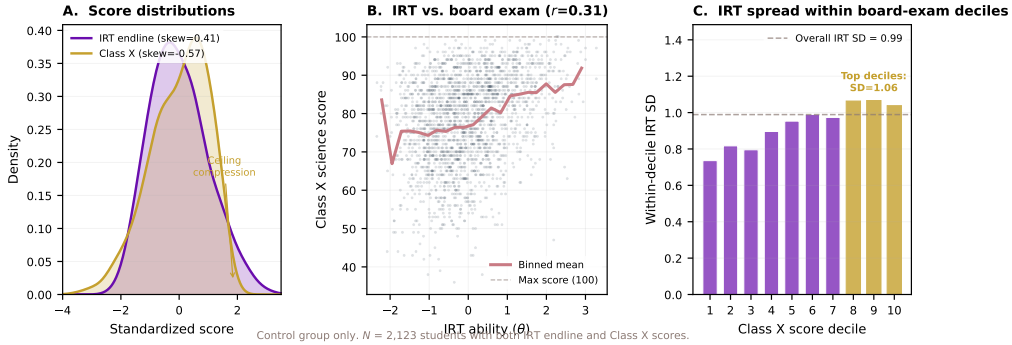


FIGURE C.6. CEILING COMPRESSION IN CLASS X BOARD EXAMS: SCIENCE

Notes: Three-panel figure for control-group students only ($N = 2{,}123$). Left panel: density comparison of IRT theta (symmetric) versus Class X z-score (left-skewed, ceiling-compressed). Center panel: scatter of IRT theta versus Class X raw score, showing fan-shape compression at the top of the board exam scale. Right panel: conditional standard deviation of IRT ability by Class X score decile, rising sharply at the top. The Pearson correlation between the IRT theta and Class X raw score is $r = 0.31$.

D. Additional Results

A. Alternative Specifications

Table D.1 presents treatment effect estimates under alternative specifications for all outcomes reported in the main tables. Column (1) reproduces the baseline specification from equation (2). Column (2) drops all controls except the treatment indicator and strata fixed effects. Column (3) adds school-level controls but omits individual-level baseline controls. Results are stable across specifications; the CPS composite ranges from 0.113 to 0.127 SD and remains significant at the 1% level throughout.

TABLE D.1—ALTERNATIVE SPECIFICATIONS: INNOVATION, HIGHER-ORDER SKILLS, AND ACADEMIC OUTCOMES

	(1) Baseline	(2) + Individual	(3) Full Specification
Composite (Table 2)			
Creative Problem-Solving (CPS)	0.116***	0.113***	0.120***
	(0.035)	(0.035)	(0.032)
Analytical Reasoning			
Fluid Intelligence	0.113*	0.111*	0.111**
	(0.057)	(0.057)	(0.050)
Problem Solving	0.109	0.108	0.142**
	(0.074)	(0.074)	(0.064)
Creativity			
Written Expression	0.038	0.032	0.040
	(0.074)	(0.073)	(0.065)
Idea Evaluation	0.104	0.103	0.100*
	(0.065)	(0.064)	(0.060)
Idea Generation	0.315***	0.312***	0.313***
	(0.083)	(0.082)	(0.077)
Socio-Emotional			
Emotional Intelligence	0.021	0.022	0.031
	(0.051)	(0.052)	(0.048)
Grit	0.020	0.018	0.025
	(0.079)	(0.079)	(0.076)
Exploration Task			
Discovery Count	0.088**	0.088**	0.094**
	(0.040)	(0.040)	(0.037)
Strategies Tried	0.071	0.070	0.077*
	(0.045)	(0.045)	(0.043)
Proximity to Optimal	−0.089**	−0.085**	−0.086**
	(0.042)	(0.042)	(0.040)
Academic Outcomes			
Mathematics	−0.022	−0.022	−0.007
	(0.078)	(0.078)	(0.067)
Science	0.018	0.017	0.027
	(0.066)	(0.065)	(0.061)
Ideas Competition (Team-Level)			
Innovation Index	0.219**	0.219**	0.227***
	(0.084)	(0.084)	(0.070)
Novelty	0.256***	0.256***	0.263***
	(0.085)	(0.085)	(0.071)
Value-Add	0.183**	0.183**	0.190***
	(0.085)	(0.085)	(0.070)
Feasibility	0.096	0.096	0.105*
	(0.069)	(0.069)	(0.060)
Willingness to Fund	0.164*	0.164*	0.178**
	(0.093)	(0.093)	(0.078)
N (Panels A–D)		4,388	
N (Panel E)		1,189	

Notes: Each cell reports the OLS treatment effect estimate with standard errors clustered at the school level (80 clusters) in parentheses. Column (1): strata fixed effects (tribal × girls-only × rural) and the baseline analog of the outcome with a missing indicator. Column (2): adds individual controls (age, household assets index, computer literacy). Column (3): the preferred specification from Tables 2–5, further adding 9 school infrastructure controls. Panel E (team-level): Columns (1) and (2) are identical because no individual-level baseline analogs or student controls are available for

B. Robustness of the CPS Composite

The CPS composite used in the main analysis is an equal-weighted average of three domain z-scores (Creativity, Analytical Reasoning, Exploration). Table D.2 shows that the main result is robust to alternative index constructions: Anderson (2008) inverse-covariance weighting, a kitchen-sink index including all available measures, and indices that sequentially drop each domain. The treatment effect ranges from 0.098 to 0.134 SD and remains significant across all specifications.

TABLE D.2—ROBUSTNESS OF THE CREATIVE PROBLEM-SOLVING INDEX

Index Definition	Outcomes	Effect (σ)	(SE)	N
3-Domain CPS (Main)	8	0.120^{***}	(0.032)	4,388
All Non-Academic (equal weight)	10	0.106^{***}	(0.028)	4,388
Individually Significant Only	7	0.131^{***}	(0.029)	4,388
Anderson (2008) ICW	10	0.102^{***}	(0.026)	4,388

Notes: Each row reports the ITT estimate on an alternative CPS index. Row 1 (main specification from Table 2): equal-weighted mean of three domain z-scores — Creativity (written expression, idea evaluation, idea generation), Analytical Reasoning (fluid intelligence, problem solving), and Exploration (discovery count, strategies tried, proximity to optimal). Each domain receives weight 1/3. Row 2: simple average of all 10 non-academic outcome z-scores, including emotional intelligence and grit. Row 3: simple average of the 7 outcomes that are individually significant ($p < 0.10$) in Tables 3–5. Row 4: inverse-covariance-weighted (ICW) index following Anderson (2008), using the control-group covariance matrix of all 10 non-academic outcomes. ICW gives more weight to outcomes carrying independent information. All specifications include strata fixed effects, baseline individual controls, 9 school infrastructure controls, and baseline outcome analogs. Standard errors clustered at the school level (80 clusters). $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

E. Heterogeneity

A. Remaining Outcomes

Table E.1 extends the heterogeneity analysis to all individual outcomes not included in the main Table 6.

TABLE E.1—HETEROGENEOUS TREATMENT EFFECTS: REMAINING OUTCOMES

	(1) Written Expression	(2) Idea Evaluation	(3) Strategies Tried	(4) Emotional Intelligence	(5) Grit	(6) Board Exams (Class X)
Panel A: By Baseline Academic Ability						
Treatment	0.001	0.022	0.061	-0.021	0.023	0.158
	(0.070)	(0.061)	(0.054)	(0.059)	(0.082)	(0.141)
$T\times$ Above Median	0.091	0.177**	0.035	0.116	0.002	-0.050
	(0.078)	(0.082)	(0.059)	(0.077)	(0.100)	(0.122)
Panel B: By Gender						
Treatment	-0.008	-0.036	0.096	0.021	0.056	0.014
	(0.093)	(0.071)	(0.071)	(0.066)	(0.093)	(0.187)
$T\times$ Girls	0.089	0.255**	-0.035	0.018	-0.058	0.228
	(0.130)	(0.122)	(0.091)	(0.095)	(0.157)	(0.230)
Panel C: By School Management (Scheduled Tribes vs. Scheduled Castes)						
Treatment	-0.113	-0.031	0.051	-0.051	-0.011	-0.011
	(0.096)	(0.089)	(0.068)	(0.066)	(0.108)	(0.157)
$T\times$ Scheduled Castes	0.289**	0.248*	0.050	0.153	0.067	0.278
	(0.133)	(0.128)	(0.087)	(0.094)	(0.154)	(0.234)
Panel D: By Location						
Treatment	0.031	0.065	0.063	0.070	0.041	0.101
	(0.124)	(0.108)	(0.067)	(0.094)	(0.127)	(0.200)
$T\times$ Rural	0.013	0.053	0.021	-0.061	-0.025	0.054
	(0.156)	(0.131)	(0.085)	(0.118)	(0.155)	(0.247)
N	4,388	4,388	4,388	4,388	4,388	5,633

Notes: This table reports heterogeneous treatment effects on outcomes that were not individually statistically significant in Tables 3–5. Outcomes are defined in Tables 3–5. Board Exams (Class X) is the total score from Class X board examinations (sum of six subjects), control-group standardized ($N = 5{,}633$). Each panel estimates an interacted regression as in Table 5. Reference groups: Panel A, below-median baseline academic ability; Panel B, boys' schools; Panel C, Scheduled Tribes (ST) schools; Panel D, urban schools. $N \approx 4{,}388$ for columns (1)–(5); $N = 5{,}633$ for column (6). Standard errors clustered at the school level in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

B. Outcome-by-Outcome Heterogeneity by Baseline Ability

Figure E.1 presents the full set of outcome-by-outcome heterogeneity estimates by baseline academic ability, complementing the summary in Table 6 Panel A.

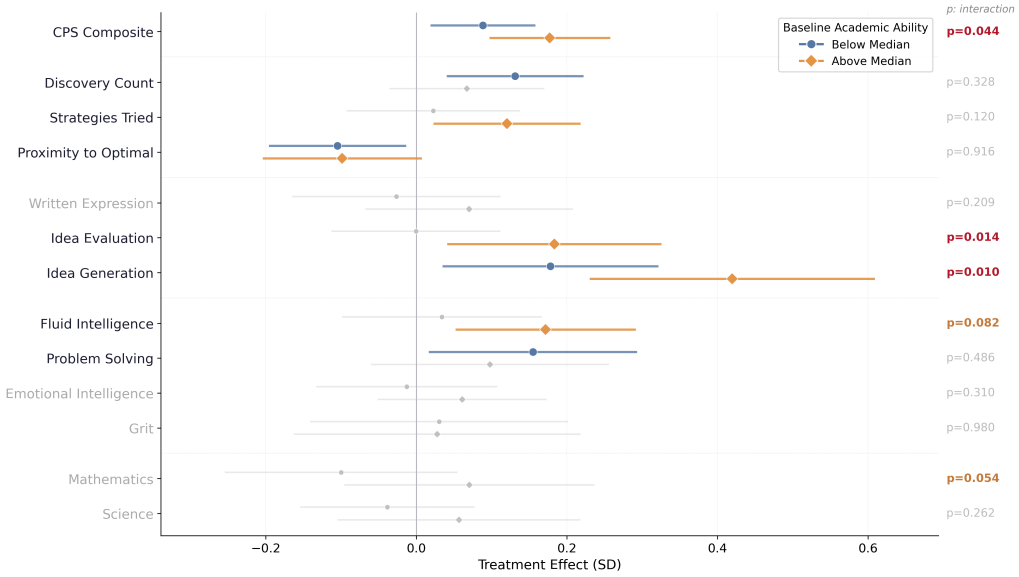


FIGURE E.1. HETEROGENEOUS TREATMENT EFFECTS BY BASELINE ACADEMIC ABILITY

Notes: Treatment effects estimated separately for below-median (blue circles) and above-median (orange diamonds) baseline academic ability via interacted regressions. Bars show 95% CIs. p-values on the right test H_0: $\beta_3 = 0$; bold indicates significance at 10%. Controls: strata FE, baseline individual controls, 9 infrastructure controls. SEs clustered at school level. $N \approx 4{,}388$.

C. Treatment Effects by Baseline Academic Quintile

Table E.2 reports treatment effects separately within each quintile of the baseline academic ability distribution, complementing the quintile figure in the main text (Figure 6).

TABLE E.2—TREATMENT EFFECTS BY BASELINE ACADEMIC QUINTILE

	Panel A: Creative Problem-Solving				*Panel B: Academic Outcomes*			
	CPS Composite	Creativity	Analytical Reasoning	Exploration	Academic Composite	Math	Science	*N*
Q1 (Lowest)	0.144***	0.106*	0.223***	0.104*	−0.018	−0.023	−0.013	878
	(0.040)	(0.061)	(0.060)	(0.054)	(0.062)	(0.070)	(0.075)	
Q2	0.115***	0.164**	0.100	0.082	−0.026	−0.108	0.056	877
	(0.043)	(0.070)	(0.060)	(0.050)	(0.065)	(0.073)	(0.074)	
Q3	0.075	0.165**	−0.001	0.061	−0.126	−0.151	−0.100	878
	(0.049)	(0.083)	(0.071)	(0.056)	(0.079)	(0.091)	(0.088)	
Q4	0.121***	0.192**	0.061	0.109**	0.052	0.025	0.080	877
	(0.046)	(0.081)	(0.076)	(0.043)	(0.083)	(0.098)	(0.085)	
Q5 (Highest)	0.258***	0.424***	0.276***	0.074	0.194**	0.218**	0.169*	878
	(0.040)	(0.069)	(0.057)	(0.051)	(0.082)	(0.085)	(0.095)	
Overall ITT	0.141***	0.208***	0.130***	0.086***	0.009	−0.011	0.029	4,388
	(0.031)	(0.053)	(0.049)	(0.031)	(0.056)	(0.065)	(0.058)	

Notes: Each cell reports the treatment effect from a separate OLS regression estimated within the indicated baseline academic ability quintile. Baseline academic ability is the average of baseline mathematics and science IRT theta scores. Panel A outcomes: CPS Composite is the equal-weighted average of three endline skill domains — Creativity (idea evaluation and idea generation), Analytical Reasoning (fluid intelligence and difficulty-weighted problem solving), and Exploration (discovery count, strategies tried, and proximity to optimal). Panel B outcomes: Academic Composite is the average of endline Mathematics and Science IRT theta scores, standardized to control-group mean and SD. All regressions include strata fixed effects, baseline individual controls, 9 school infrastructure controls, and baseline academic ability. Standard errors clustered at the school level in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

D. Robustness of Academic Complementarity

Table E.3 documents the robustness of the positive treatment effect on academic outcomes for students in the top quintile of the baseline ability distribution. Panel A shows that the coefficient is positive (+0.15 to +0.20 SD) across all specifications, from a bare treatment-only regression to the full KLK specification. Panel B demonstrates that the result is not an artifact of the quintile cutpoint: positive effects appear for the top tercile (+0.155 SD, $p = 0.056$), top quartile (+0.192 SD, $p = 0.024$), and top quintile (+0.194 SD, $p = 0.021$). Panel C reports a full-sample interacted regression, where the treatment-by-top-quintile interaction is significant for the academic composite ($p = 0.019$) and for mathematics specifically ($p = 0.003$). Panel D provides the randomization inference p-values ($p = 0.063$ for the composite, $p = 0.046$ for mathematics) and shows that differential attrition in the top quintile works *against* a positive finding: treatment students are 3.9 percentage points less likely to attrit than control students.

TABLE E.3—ROBUSTNESS OF ACADEMIC COMPLEMENTARITY AT TOP OF ABILITY DISTRIBUTION

	Academic Composite	Mathematics	Science	*N*
Panel A: Specification Robustness (Top Quintile Subsample)				
Treat only (no controls)	0.151	0.190	0.112	878
	(0.115)	(0.117)	(0.131)	
Strata FE	0.160	0.204*	0.115	878
	(0.103)	(0.103)	(0.121)	
+ Individual controls	0.160	0.203*	0.117	878
	(0.103)	(0.102)	(0.121)	
+ School infrastructure	0.198**	0.223**	0.174*	878
	(0.087)	(0.087)	(0.101)	
Full KLK (main spec)	0.194**	0.218**	0.169*	878
	(0.082)	(0.085)	(0.095)	
Panel B: Alternative Cutpoints (Full KLK Specification)				
Top tercile (top 33%)	0.155*	0.158*	0.152*	1,463
	(0.080)	(0.084)	(0.088)	
Top quartile (top 25%)	0.192**	0.218**	0.166*	1,097
	(0.083)	(0.083)	(0.096)	
Top quintile (top 20%)	0.194**	0.218**	0.169*	878
	(0.082)	(0.085)	(0.095)	
Panel C: Full-Sample Interaction Test (Treat × Top Quintile)				
Treatment ($\hat{\beta}_1$, non-Q5)	−0.033	−0.068	0.003	4,388
	(0.055)	(0.067)	(0.055)	
Treat × Top Quintile ($\hat{\beta}_3$)	0.199**	0.276***	0.122	
	(0.083)	(0.089)	(0.098)	
Treatment for Q5 ($\hat{\beta}_1 + \hat{\beta}_3$)	0.166*	0.207**	0.124	
	(0.092)	(0.093)	(0.108)	
Panel D: Randomization Inference and Sample Selection				
RI *p*-value (999 permutations)	0.063	0.046		
Differential attrition (T − C)		−0.039 ($p = 0.102$)		
Treatment attrition rate		0.182		
Control attrition rate		0.221		

Notes: This table documents the robustness of the positive treatment effect on academic outcomes for students in the top quintile of the baseline academic ability distribution. Panel A varies the control set within the top quintile subsample ($N = 878$). Panel B varies the definition of "top" of the ability distribution using the full KLK specification. Panel C reports a full-sample interacted regression: $y_i = \beta_0 + \beta_1 T_j + \beta_2 TopQ5_i + \beta_3 (T_j \times TopQ5_i) + X_i'\theta + \varepsilon_i$. Panel D reports the randomization inference *p*-value (999 within-strata school-level permutations) and differential attrition rates between treatment and control for the top quintile. The Academic Composite is the average of endline Mathematics and Science IRT theta scores, control-group standardized. All Panel A–B regressions include the controls listed in the first column. Standard errors clustered at the school level in parentheses.

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

E. CPS Profiles by Sub-Component

Figure E.2 replicates the talent profiles analysis separately for each of the three CPS sub-components (Creativity, Analytical Reasoning, Exploration).

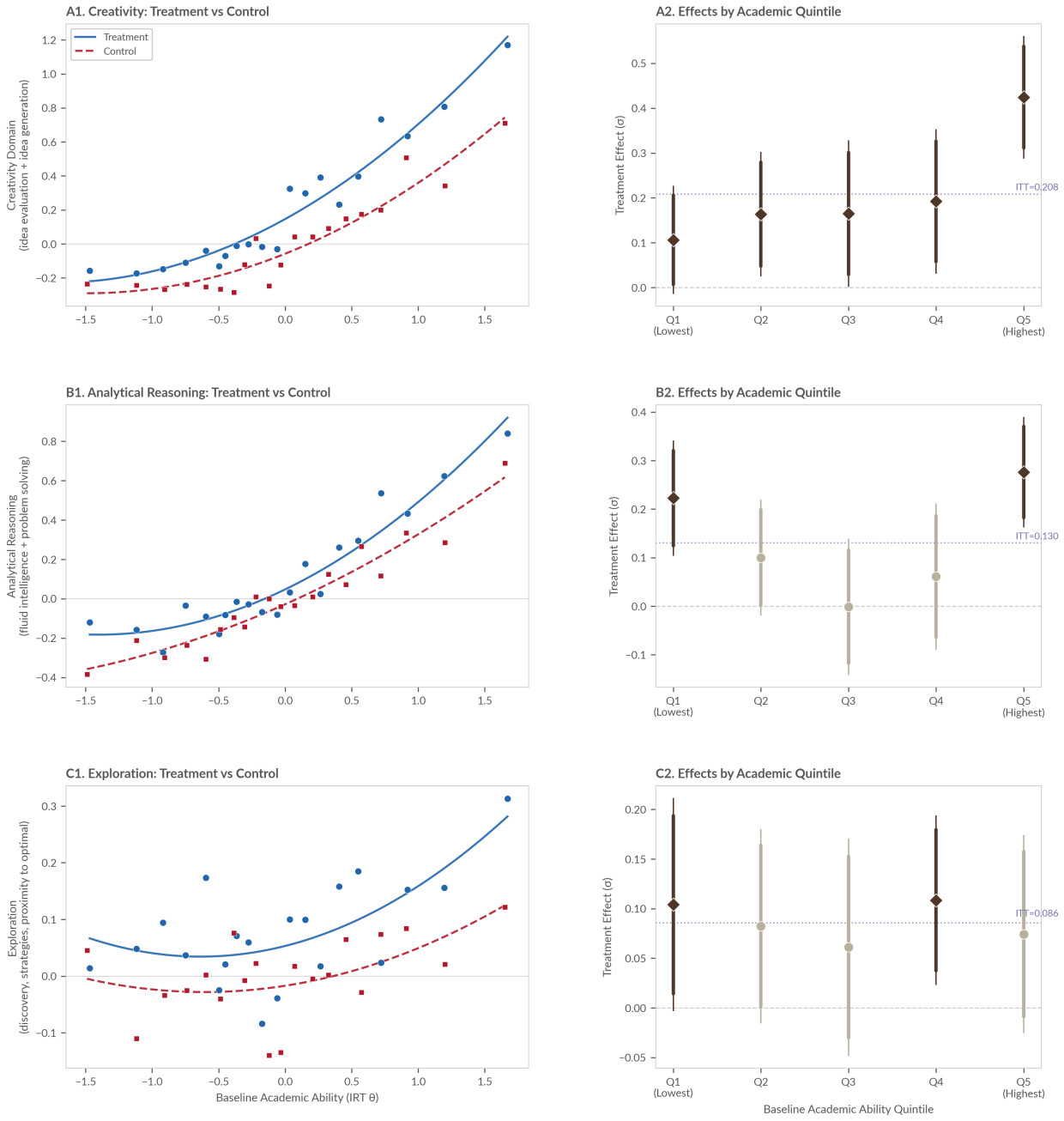


FIGURE E.2. CPS PROFILES BY SUB-COMPONENT

Notes: Rows show Creativity (A), Analytical Reasoning (B), and Exploration (C) domains. Left panels: binned scatterplots of endline domain score against baseline academic ability. Right panels: quintile treatment effects. See Figure 5 notes for specification details.

F. Heterogeneity by Gender, School Management, and Location

Figures E.3, E.4, and E.5 present the full set of outcome-by-outcome heterogeneity estimates by gender, school management type, and location, complementing the summary in Table 6 Panels B–D.

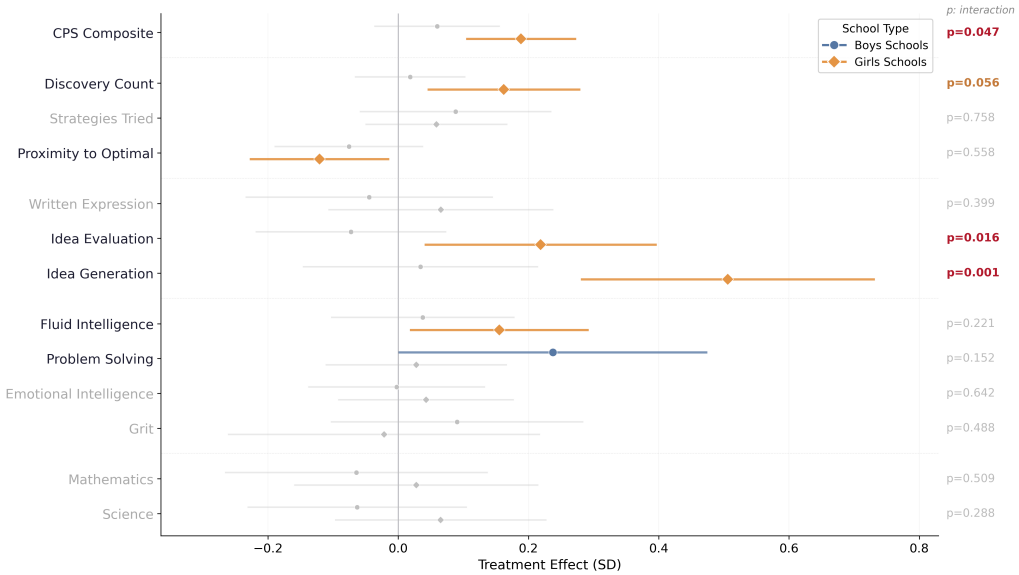


FIGURE E.3. HETEROGENEOUS TREATMENT EFFECTS BY GENDER

Notes: Treatment effects for boys' schools (blue) and girls' schools (orange) from interacted regressions. Bars show 95% CIs. p-values test the interaction; bold = significant at 10%. Controls: strata FE (gender removed, entered directly), baseline controls, infrastructure. SEs clustered at school level. $N \approx 4{,}388$.

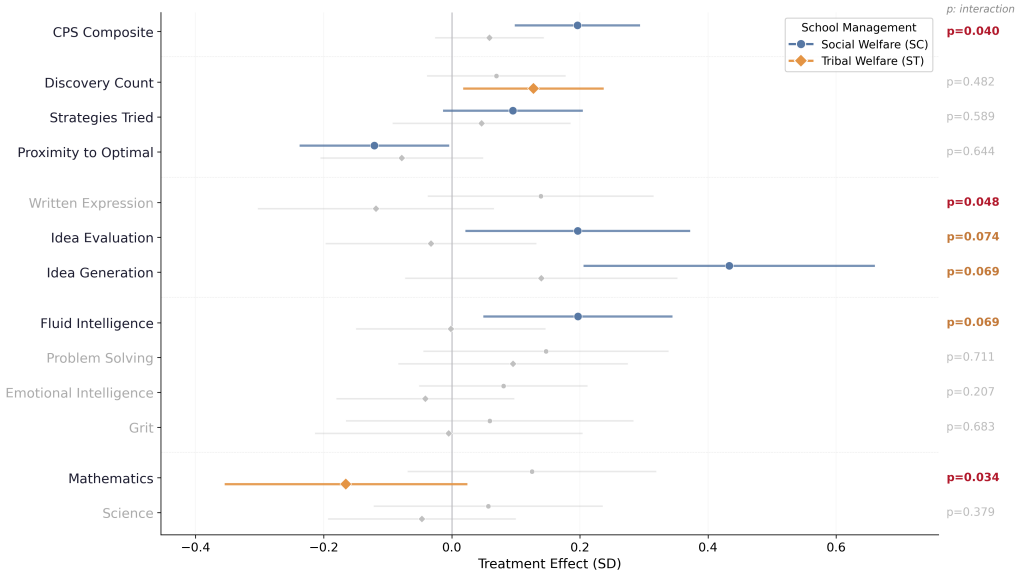


FIGURE E.4. HETEROGENEOUS TREATMENT EFFECTS BY SCHOOL MANAGEMENT

Notes: Treatment effects for Tribal Welfare (ST) and Social Welfare (SC) schools from interacted regressions. Bars show 95% CIs. p-values test the interaction; bold = significant at 10%. Controls: strata FE (tribal removed, entered directly), baseline controls, infrastructure. SEs clustered at school level. $N \approx 4{,}388$.

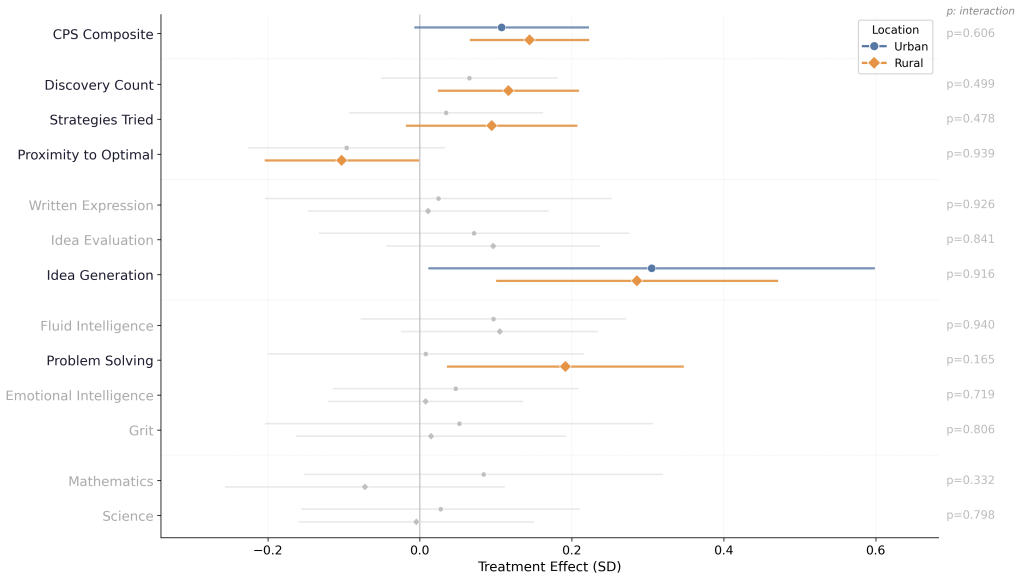


FIGURE E.5. HETEROGENEOUS TREATMENT EFFECTS BY LOCATION

Notes: Treatment effects for urban (blue) and rural (orange) schools from interacted regressions. Bars show 95% CIs. *p*-values test the interaction; bold = significant at 10%. Controls: strata FE (rural removed, entered directly), baseline controls, infrastructure. SEs clustered at school level. $N \approx 4{,}388$.

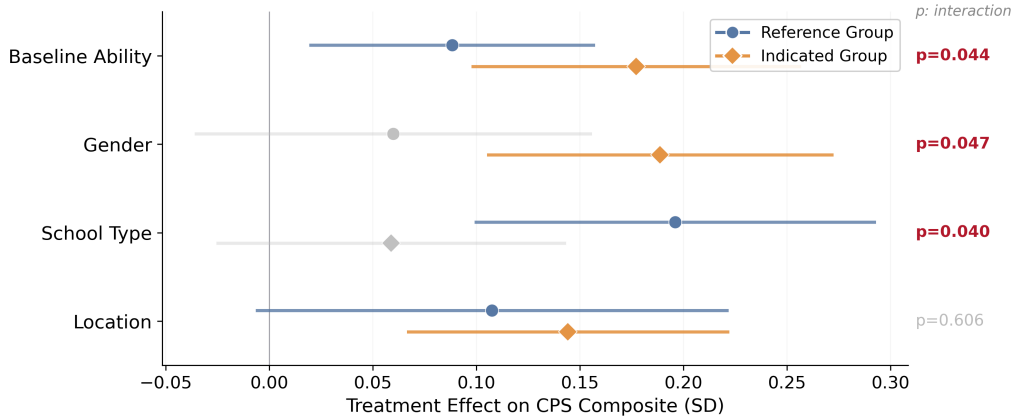


FIGURE E.6. CPS COMPOSITE: HETEROGENEOUS TREATMENT EFFECTS ACROSS ALL DIMENSIONS

Notes: Treatment effects on the CPS composite for reference (blue circles) and indicated (orange diamonds) groups across four heterogeneity dimensions: baseline academic ability (above/below median), gender (boys'/girls' schools), school management (ST/SC), and location (urban/rural). Bars show 95% CIs. *p*-values test the interaction; bold = significant at 10%. All specifications use the interacted KLK regression from Table 6. SEs clustered at school level. $N \approx 4{,}388$.

G. Innovation Sector and Beneficiary Distributions

Figures E.7–E.10 present the full sector choice and target beneficiary distributions across treatment status, gender, location, and school type.

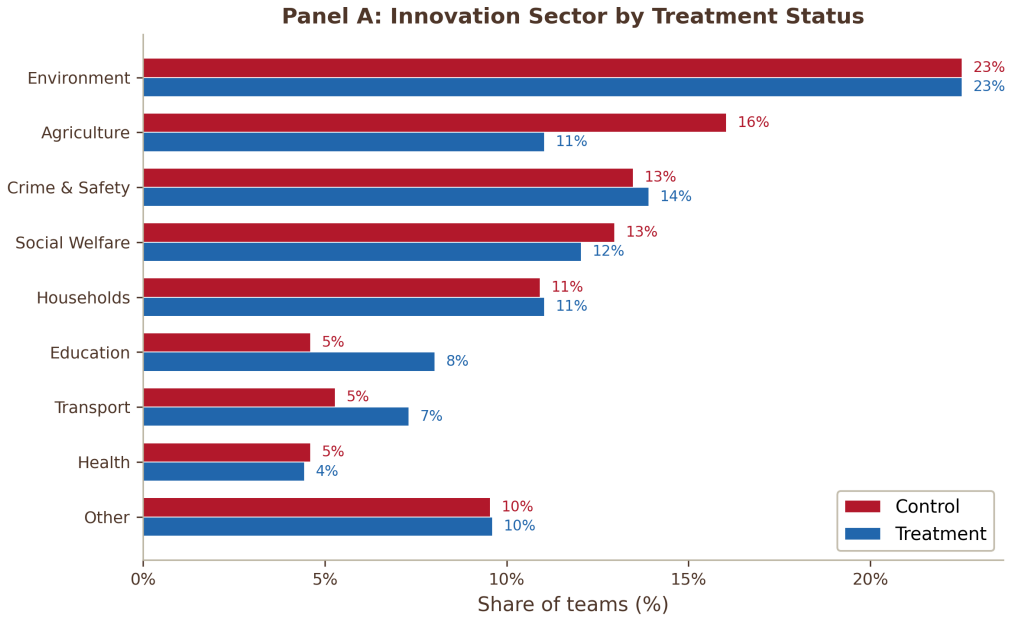


FIGURE E.7. INNOVATION SECTOR CHOICE BY TREATMENT STATUS

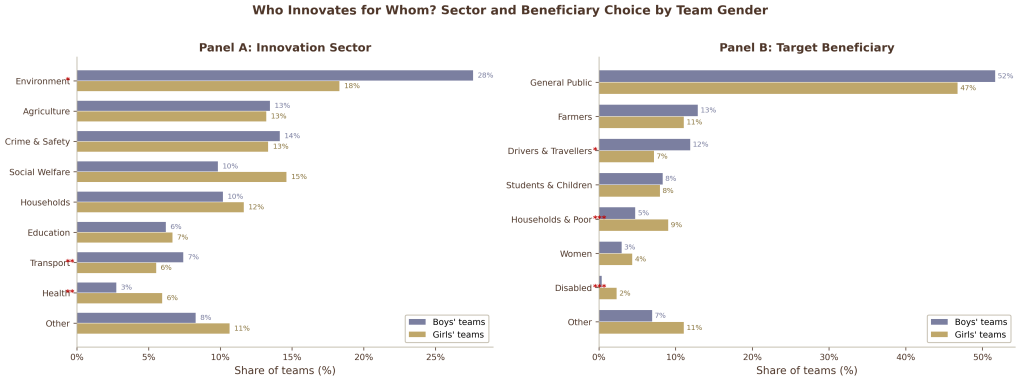


FIGURE E.8. INNOVATION SECTOR AND TARGET BENEFICIARY BY TEAM GENDER

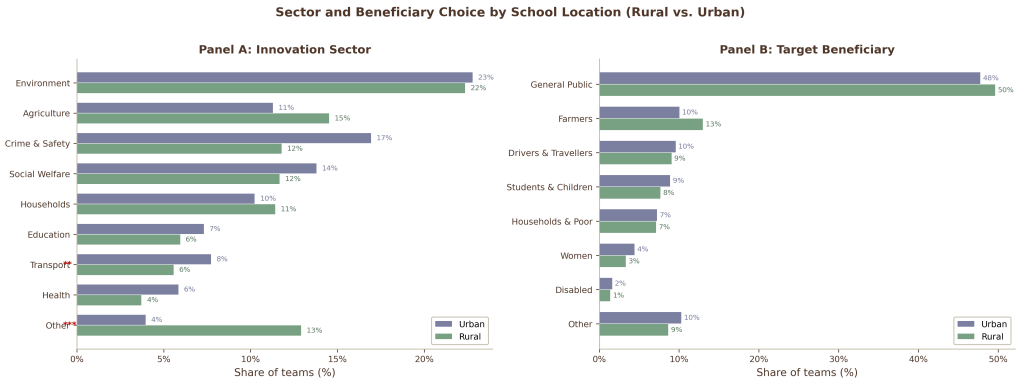


FIGURE E.9. INNOVATION SECTOR AND TARGET BENEFICIARY BY SCHOOL LOCATION

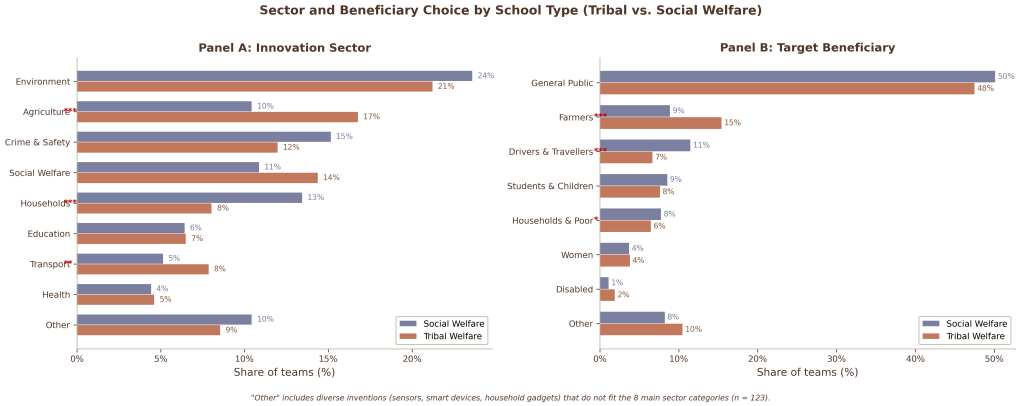


FIGURE E.10. INNOVATION SECTOR AND TARGET BENEFICIARY BY SCHOOL TYPE

F. Pre-Registration and Transparency

This appendix provides a systematic mapping between the pre-analysis plan registered at the AEA RCT Registry (AEARCTR-0009696, initially registered August 8, 2022; last updated May 26, 2025) and the analyses reported in this paper. The registration includes three supporting documents uploaded in January 2024 containing the complete assessment instruments.

A. Pre-Registered Primary Outcomes

Table F.1 maps each pre-registered outcome to its status in the paper. All primary outcomes specified in the registry are reported; four are redirected to a companion paper or replaced by performance-based alternatives, as noted in Section A.

TABLE F.1—PRE-ANALYSIS PLAN: OUTCOME MAPPING

PAP Outcome	**Paper Outcome**	**Status**	**Notes**
Team Outcomes (Ideas Competition)			
Innovation Scale Score	Innovation Index	Reported	Composite of novelty, feasibility, value-add (Table 3)
Willingness to Fund	Willingness to Fund	Reported	Table 3
Funding Raised	Funding Competition	Reported	Figure 4; subsample ($N = 78$ teams)
Community Feedback	—	Not reported	Not collected at endline
Individual Outcomes			
ICAR Matrix Reasoning	Fluid Intelligence	Reported	Table 4
Math & Science (Grade 7)	Academic Composite	Reported	Table 2; IRT-scored
Big 5 Personality Traits	—	Dropped	Collected at midline only; dropped at endline to reduce test fatigue
Innovators DNA Survey	—	Not collected	Replaced by exploration task before baseline
Risk-Taking (Gneezy–Potters)	—	Dropped	No variation at midline; dropped at endline to reduce test fatigue
Lemonade Test (Ederer–Manso)	Exploration Task	Reported	Adapted as multi-armed bandit; Table 3, Figure C.1
Social Skills (Mind in Eyes)	Emotional Intelligence	Reported	Table 4
Social Network	—	Companion	Reserved for companion paper on team composition
Outcomes Added to Assessment Instruments (Jan 2024)			
Creativity (PISA-adapted)	Written Expression, Idea Evaluation, Idea Generation	Reported	Three domains; Table 4
Problem Solving (Alan et al.)	Problem Solving	Reported	Table 4
Grit (Alan et al.)	Grit	Reported	Table 4
Secondary / Added Post-Registration			
Board Exam (Class X)	8 board exam subjects	Reported	Administrative data; Table 5. IRB approved at Stanford (Apr 2025)

B. Pre-Registered Secondary Analyses

The registry specifies two secondary analyses: (i) the effect of team composition (homogeneous vs. heterogeneous teams based on social skills) on innovation outcomes, and (ii) heterogeneity analysis examining how individual traits predict team success. The team composition analysis is reserved for a companion paper, as teams were separately randomized within both treatment and control schools.

C. Pre-Specified vs. Exploratory Analyses

Table F.2 classifies the heterogeneity and supplementary analyses reported in this paper by their pre-registration status.

TABLE F.2—CLASSIFICATION OF ANALYSES: PRE-SPECIFIED VS. EXPLORATORY

Analysis	**Classification**	**Basis**
Main treatment effects (CPS, exploration, innovation, skills, academics)	Pre-specified	All outcomes in registry or Jan 2024 instruments
CPS Composite index	Post-hoc	Equal-weighted mean of pre-registered outcomes; robustness to ICW shown in Table D.2
Gender heterogeneity	Pre-specified	Stratification variable; registry specifies heterogeneity analysis
Quintile decomposition (academic ability)	Exploratory	Not in registry; motivated by board exam patterns. Formal interaction test reported
Hidden Innovators (talent profiles)	Exploratory	Profile framework developed ex post; labeled as exploratory in text
Sector choice / entrepreneurial intentions	Exploratory	Not in registry; descriptive analysis of downstream career intentions
Board exam effects	Pre-specified	Listed as secondary outcome; data collection added with separate IRB

D. Deviations from the Pre-Analysis Plan

The following substantive deviations from the original registration are noted:

1. **Endline battery revision.** The Innovators DNA survey was never administered; it was replaced by the exploration task before baseline data collection. The risk-taking task (Gneezy and Potters, 1997) and Big Five personality inventory were administered at midline (Year 1) but dropped from the endline battery—the risk-taking task showed no variation and Big Five is self-reported—to reduce test fatigue and make room for three new performance-based assessments: creativity (PISA-adapted), problem solving (Alan et al., 2019), and grit (Alan et al., 2019).

2. **Creativity assessments.** The three PISA-adapted creativity measures (written expression, idea evaluation, idea generation) were not in the original August 2022 registration but were included in the assessment instruments uploaded to the registry in January 2024, before endline data collection began.

3. **Exploration task adaptation.** The "Lemonade Test" referenced in the registry was substantially adapted into a multi-armed bandit game with a different payoff structure, five decision dimensions, and tablet-based implementation. The core construct—experimentation and tinkering behavior—is preserved from the original Ederer and Manso (2013) design.

4. **CPS composite.** The headline summary index was not pre-registered as a named outcome. It is an equal-weighted average of all pre-registered individual skill z-scores, constructed following Anderson (2008). Table D.2 demonstrates robustness to inverse-covariance weighting and alternative variable inclusion.

5. **Companion paper.** Social network data and team composition effects (homogeneous vs. heterogeneous teams) are reserved for a separate paper, as the within-school team randomization constitutes an independent experiment.

E. Year 1 Midline Analysis

The Year 1 midline assessment was administered approximately six months after baseline (February–March 2023), during the first full academic year following COVID-related school closures. Students in these residential schools were readjusting to structured schooling after extended disruption. Table F.3 reports treatment effects on three pre-registered midline outcomes. Mathematics shows a significant negative effect (-0.137 SD, $p = 0.021$ by randomization inference). Science is null (-0.006 SD). Emotional intelligence (RMET) shows a marginally significant negative effect (-0.101 SD, $p = 0.097$). These outcomes were pre-registered and are reported for transparency.

The dip is fully transient. Table F.4 presents treatment-control differences across three waves for a balanced panel of 2,884 students with non-imputed data at all time points. Mathematics follows a V-shaped trajectory: the treatment-control gap moves from -0.021 SD at baseline to -0.148 SD at midline, then reverses to $+0.014$ SD at endline and $+0.038$ SD on Class X board examinations administered one year after the intervention. Within-student changes in mathematics are -0.127 SD from baseline to midline ($p = 0.001$) and $+0.162$ SD from midline to endline ($p < 0.001$), with a net change of $+0.035$ SD (not significant). RMET shows the same pattern: -0.165 SD at midline, recovering to $+0.016$ SD at endline. If students had genuinely lost mathematical knowledge or cognitive skills, the deficit would not reverse completely—and would not be absent on externally administered, high-stakes state board examinations one year later.

The midline survey was administered on tablets with three major content blocks—academic assessment, emotional intelligence (RMET), and an incentivized exploration game—presented in randomized order across students. For completeness, Table F.5 examines potential mechanisms for the mathematics dip. Of twelve hypotheses investigated, eight are rejected outright, one (effort reduction) is partially supported, one (time of day) is inconclusive, and one (surveyor identity) could not be tested due to data limitations. The dip is not ability loss (both arms forget approximately 45 percent of previously correct items; the treatment differential is only 2.8 percentage points). It is not survey fatigue (the gap is larger, not smaller, when the academic section appears early in the survey). It is not driven by specific schools (the gap is diffuse across 23 of 40 treatment schools and does not persist to endline). It is not differential response quality (response entropy and straight-lining rates are identical across arms). The most consistent interpretation is motivation displacement: treatment students who had just completed the exploration game invested less effort on routine academic items. The gap is driven by the easiest items, where effort matters most; by the fastest students, who are most disengaged; and is largest when the academic section immediately follows the game. Science shows zero treatment effect at every survey position. The post-COVID context likely amplified this effect, as students were returning to structured assessment after extended school closures.

The midline dip is uniform across the baseline ability distribution. In an interacted specification splitting students at the baseline academic median, the treatment-by-above-median interaction on mathematics is 0.042 SD ($p = 0.569$). A tercile decomposition shows the middle tercile is hit hardest (-0.226 SD), not the bottom (-0.102 SD)—inconsistent with any ability-based mechanism. In the three-wave panel, above-median students recover more strongly: they end endline at $+0.106$ SD, while below-median students partially recover to -0.070 SD. The uniformity of the dip across ability levels is additional evidence for the measurement artifact interpretation: if this were real learning loss, it would concentrate among weaker students.

TABLE F.3—YEAR 1 MIDLINE TREATMENT EFFECTS

	Mathematics (1)	Science (2)	Emotional Intelligence (RMET) (3)
Treatment	-0.137***	-0.006	-0.101*
	(0.049)	(0.047)	(0.051)
	[0.021]	[0.916]	[0.097]
Control mean	0.000	0.000	-0.000
Control SD	1.000	1.000	1.000
Observations	4,620	4,620	4,620
R^2	0.233	0.261	0.137

Notes: Standard errors clustered at school level (80 clusters) in parentheses. Randomization inference p-values (999 permutations) in square brackets. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$ All outcomes are z-scored to the control group (mean = 0, SD = 1). Midline data collected approximately 6 months after baseline (Feb–Mar 2023), during Year 1 of the intervention. Controls: strata fixed effects (tribal × girls × rural), individual controls (age, SES index, computer skills), school infrastructure, and baseline outcome with missing indicator.

TABLE F.4—YEAR 1 MIDLINE DIP AND RECOVERY: THREE-WAVE ACADEMIC TRAJECTORY

	Baseline (1)	Midline (2)	Endline (3)	Class X (4)
Mathematics				
Treatment − Control	-0.021	-0.148***	+0.014	+0.038
	(0.576)	(0.000)	(0.708)	(0.328)
Science				
Treatment − Control	-0.009	-0.051	+0.032	-0.061
	(0.810)	(0.179)	(0.390)	(0.103)
RMET				
Treatment − Control	—	-0.165***	+0.016	—
		(0.000)	(0.660)	
Observations	2,884 students with baseline + midline + endline data			

Notes: Balanced panel of students with non-imputed baseline, midline, and endline assessment data. Each cell reports the treatment–control difference in standard deviations (z-scored to the control group at each wave). p-values from two-sample t-tests in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$ Class X columns use total marks (internal + external) from administrative board exam records. The midline dip in Mathematics and RMET fully reverses by endline, consistent with a transient survey context effect rather than ability loss.

TABLE F.5—MIDLINE MATH LOSS: MECHANISM INVESTIGATION

	Hypothesis	Verdict	Key Evidence
H1	Ability loss (forgetting)	Rejected	Both arms forget ∼45%; differential 2.8 pp. Full EL recovery.
H2	Effort / rushing	Partial	T 27s faster (p=.006), gap largest in fastest quartile (-0.289 SD). Time control attenuates by 6.8%.
H3	Survey fatigue (block order)	Rejected	Gap *larger* when academics early (-0.177 SD) vs. late (-0.098 SD).
H4	Motivation displacement	**Supported**	Best fit: gap largest post-game, full EL recovery, driven by easy items.
H5	Bad schools / surveyors	Rejected	Diffuse (23/40 T schools), does not persist to endline.
H6	Item-specific forgetting	Rejected	Gap similar on shared (-0.138 SD) and new (-0.153 SD) items.
H7	DIF (item interpretation)	Rejected	IRT parameters similar across arms. Gap largest on easiest items.
H9	Straight-lining	Rejected	Response entropy, run length, extreme scores identical across arms.
H10	Within-section fade	Rejected	No gradient early→late within math block.
H11	Game perf. × math	Rejected	Interaction $\beta \approx 0$ (p=.94). Uniform across profit terciles.
H12	Time of day	Inconclusive	Gap absent 12–14h, present 10–12h and 14–17h. No clean pattern.
H13	Het by BL ability	Rejected	Interaction p=.569 (math). Gap uniform across median split and terciles.

Notes: Summary of mechanism tests for the Year 1 midline math treatment effect (-0.175 SD, $p < 0.001$). H8 (surveyor identity) not testable — surveyor IDs unavailable. The weight of evidence supports H4 (motivation displacement): treatment students, having completed an engaging creative task, invested less effort on routine cognitive items. The effect is transient and fully reverses by endline (Table F.4).

G. Sample Student Ideas

This appendix showcases a selection of student-led innovations developed by treatment-group teams during the two-year school-based innovation training program. Projects span diverse domains—from financial inclusion and public infrastructure to mobility, hygiene, and environmental sustainability—and illustrate the range of problems that students independently identified and attempted to solve.

Everyday Access and Inclusion

Innovation Title	**Idea Description**
Smart Bank Form Assistant	A voice-assisted machine that helps illiterate or elderly users fill out bank forms independently. The system prompts users in their preferred language and prints completed forms for submission. Designed to reduce dependency and increase financial access.
ATM for Torn Note Exchange	An ATM add-on that verifies and replaces torn currency notes, allowing users to exchange damaged bills without visiting a bank. Addresses friction in everyday financial transactions.
Smart Message Identifier	A phone-based system that flags spam or scam messages in red for non-literate users, helping prevent digital fraud. Uses simple color cues to guide decision-making and protect vulnerable users.

Public Infrastructure and Safety

Innovation Title	**Idea Description**
Smart Drainage Waste Management	A three-part system to filter, compress, and recycle waste from open drainage lines. Designed to reduce blockages, improve sanitation, and streamline municipal operations.
Smart Weight Monitoring for Bridges	A sensor-based alert system to monitor vehicle weight on bridges and prevent overloading. Includes visual warnings and links to traffic management systems.
Delivery Vehicle with Safety Features	A redesigned delivery vehicle equipped with a roof, food compartments, and airbag system to protect gig workers from accidents and weather conditions.

Health, Hygiene, and Daily Life

Innovation Title	**Idea Description**
Smart Hygiene Toilet Lock	A toilet door lock that only opens after water is poured, enforced via sensor and buzzer. Designed to encourage hygienic habits in school washrooms.
Smart Pill-Reminder Water Bottle	A smart water bottle with pill compartments and built-in reminders. Combines hydration and medication adherence for elderly or chronically ill users.
Smart Vegetable Cleaner & Cutter	A three-compartment machine for washing, testing, and cutting vegetables efficiently. Designed for food safety in households with busy caregivers.

Mobility and Accessibility

Innovation Title	Idea Description
Smart Walking Stick with Rolling Bag	A bag attachment with wheels connected to a walking stick, allowing elderly users to carry loads hands-free while maintaining balance.
Height-Adjustable Classroom Benches	Adjustable school benches that improve posture and comfort by accommodating different student heights and seating preferences.
Height-Adjustable Walker	A mobility aid with front/back height control for safe use on stairs. Designed for people with disabilities or post-surgery needs.

Learning and Motivation

Innovation Title	Idea Description
Smart Wristband for Gym Ranker	A wearable device that tracks gym performance and displays student rankings on a shared screen to motivate fitness through peer comparison.
FunFit Kids Gym	A gym designed specifically for children, using colorful equipment and interactive challenges to make physical activity more engaging.
Smart Mini Playground with Rewards	A mini playground system with hurdle sensors and a reward wheel to encourage active play and goal-setting among children.

H. Qualitative Analysis of Student Feedback

To complement the quantitative findings, I collected open-ended feedback from treated students at endline, asking them to describe their experience with the Inqui-Lab program. This yielded 2,524 text entries, of which 1,746 contained substantive responses (text length > 50 characters). I analyzed this corpus using an LLM to classify each response by sentiment and to identify recurring themes.

Sentiment analysis. The LLM classified 97 percent of substantive responses as positive, 3 percent as neutral, and 0 percent as negative. Five dominant themes emerged from the thematic coding: problem-solving confidence, community responsibility, positive change, sense of possibility, and career aspirations. Representative quotes from each theme are presented below.

Problem-Solving Confidence.

"I liked the Think and Make program from the Inqui-Lab Foundation because this is highly beneficial for the country's development. Additionally, children can learn to take more initiative in problem-solving. As a student leader, I gained valuable experience in teaching problem-solving methods. This program should achieve even greater success."

This theme aligns with the quantitative findings on improved creative problem-solving (0.19 SD). Students frequently mentioned enhanced confidence in approaching complex problems and applying structured thinking processes.

Community Responsibility.

"I have learned a lot from the Think and Make program. This is so interesting and exciting. In this Think and Make program, we have created 2 prototypes. I have also learned how to solve the surrounding problems and how to overcome them. . . "

Students consistently referenced applying innovation skills to address community challenges, reflecting the pattern observed in the Ideas Competition data where students directed attention to locally relevant problems.

Positive Change.

"I am very happy with this Inqui Lab program in 8th class. I feel delighted as it has improved my thinking abilities significantly. Many activities are taking place, and I am excited about the positive changes occurring in our school. We are advocating against plastic usage and encouraging the use of steel bottles. Similarly, we are promoting water conservation, preventing food wastage, and emphasizing the importance of saving plants. I am extremely happy about all of the changes happening around me."

This theme emphasizes the tangible changes students implemented in their immediate environments, providing evidence that innovation training translates into real-world actions beyond classroom activities.

Sense of Possibility.

"I liked this program because I think this is very innovative, and am interested in this program. I solved so many problems through this Think and Make program. This is really useful for our life, and I learn if we have any problem in our areas or surroundings, we can easily solve the problem, so I thank to Think and Make program."

Many students expressed a newfound sense of agency and possibility, suggesting that the program may foster self-efficacy and a growth mindset—attributes that complement the quantitative evidence on exploration behavior and creative problem-solving.

Career Aspiration.

"I will find some problems in our house, gurukulam, school in my surroundings, and then find their solutions.

Inqui-lab's Think and Make program make me so happy. I will become a scientist in my life thank-you for this giving opportunity sirs and madams."

This theme connects to the quantitative findings on career aspirations, particularly the increase in teacher career aspirations observed in the main results. Students articulated specific future career goals related to innovation, problem-solving, and scientific pursuits.

The consistent patterns across hundreds of independent responses provide complementary evidence for the mechanisms through which the innovation curriculum influenced student outcomes. The emphasis on problem-solving, community relevance, and an expanded sense of possibility aligns with the theoretical framework of skill development discussed in the main text.