# Can Innovation Be Taught? Evidence from a School-Based RCT in India

*Saloni Gupta*[†]

Working Paper

April 4, 2025

Updated Frequently. Click here for the latest version


## Abstract

Innovation fuels long-run economic growth, yet education systems in developing countries often overlook the skills required for innovation. This paper provides the first experimental evidence that students can learn core innovation-related skills. I conduct a large-scale clustered randomized controlled trial with 4,800 eighth-grade students in India, comparing a two-year school-based innovation training program to a control group with access to self-directed project opportunities. Because existing measures of innovation are scarce and rarely validated in low-income school settings, I develop a novel, multi-method toolkit that includes a strategic exploration task, expert-rated idea quality assessments, and a live funding competition. Treated students significantly outperformed controls across all three domains (0.10–0.24 SD) and secured 24 percentage points more external funding. The program also improved fluid intelligence and problem-solving, though these remained distinct from innovation outcomes. Heterogeneity analyses show complementary gains in innovation and math for students with above-median baseline scores (0.16–0.20 SD), and gains in higher-order skills but math declines (–0.14 SD) for those below the median. These patterns suggest that lower-performing students reallocated effort away from abstract math toward more applied, hands-on tasks where they experienced greater success. The results offer a proof of concept: innovation skills are malleable, and marginalized students can develop them—reframing assumptions about when and where valuable ideas can emerge.


*JEL Classification: I21, I25, O31, J24, C93 Keywords: Educational Innovation, Human Capital Development, Experimental Economics, Skill Formation, Educational Interventions*

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. I am especially grateful to myself for staying committed to this project over six years, despite significant challenges. In the first year of data collection, I worked two jobs, wrote multiple grant applications, and ran field operations without funding for the midline or a PhD stipend. The project received institutional support only in its second year, through the NAEd/Spencer Dissertation Fellowship and the Weiss Fund. I remained committed despite skepticism from parts of the economics community about the feasibility of measuring innovation among students, and about the importance of studying higher-order skills before foundational skills are mastered by disadvantaged students in India. This paper is a reflection of that conviction—and of what is possible when research questions are grounded in purpose and pursued with care. The RCT is registered with the AEA Trial Registry (RCT ID: AEARCTR-0009696).

[†]salonig@stanford.edu, Stanford King Center on Global Development, Stanford University

## 1. Introduction

Innovation drives long-run economic growth through the generation of new ideas. Yet the capacity to innovate remains highly unequally distributed across populations. Children from disadvantaged backgrounds, especially in developing countries, are less likely to become inventors not necessarily because of lower ability, but because of differential exposure, incentives, and institutional complements (Bell et al., 2019; Aghion et al., 2017; Jones, 2022b). While economists have documented how early environmental factors shape who becomes an innovator, a fundamental question remains: Can the human capital component of the innovation production function be systematically developed through educational interventions?

This question is increasingly relevant as numerous countries have introduced school-based innovation programs over the past decade.[1] These initiatives diverge from traditional STEM instruction by emphasizing exploration, analogical reasoning, and iterative problem-solving rather than content mastery alone. They aim to develop the behavioral foundations of innovation—the cognitive and non-cognitive skills that enable individuals to identify problems, generate novel solutions, and refine ideas through feedback. This paper investigates whether such innovation-related skills are malleable during adolescence, and whether structured, scalable school-based interventions can effectively cultivate them.

If innovation skills can be developed in adolescence, the economic implications extend across both the technological frontier and developing economies. At the frontier, broadening the innovator pipeline directly diversifies who becomes inventor and which problems receive attention (Bell et al., 2019; Aghion et al., 2017; Einiö et al., 2023). In developing economies, where 88% of workers operate in small or informal enterprises (ILO, 2018), innovation capacity transforms productivity. Firm-level evidence demonstrates that interventions enabling workers to identify process inefficiencies yield productivity gains of 8-30% (Bloom et al., 2013; Jensen and Miller, 2018; Adhvaryu et al., 2018; Atkin et al., 2017). While these studies target adults with direct training, school-based programs offer a more scalable approach by embedding these skills during formative years.

This paper establishes whether structured educational interventions develop the cognitive foundations of innovation among adolescents. I evaluate a two-year school-based innovation training program implemented in government residential schools in Telangana, India. The study was conducted as a clustered randomized controlled trial across 80 schools, with 40 assigned to treatment and 40 to control. These schools serve exclusively students from constitutionally protected marginalized communities (Scheduled Castes and Scheduled Tribes), providing a setting to examine whether innovation-related behaviors can be developed among populations historically excluded from formal inno-

[1] Examples include India's Atal Tinkering Labs and Delhi's Entrepreneurship Mindset Curriculum; Rwanda's ICT and Innovation Policy (2015); Thailand's Education 4.0 framework; and several design-focused reforms across Latin America (NITI Aayog, 2020; Delhi Government, 2019; Ministry of ICT and Innovation, Rwanda, 2019; Ministry of Education, Thailand, 2018; Blikstein et al., 2022).

vation pipelines. I follow approximately 5,000 students from Grade 8 through Grade 9 during academic years 2022–2024.

A critical challenge in studying innovation-related skills is measurement. While most educational interventions are evaluated using test scores or behavioral proxies of engagement, there are few validated tools for capturing exploratory behavior and idea generation in school settings—particularly in low-resource environments. This study develops a multi-method measurement strategy grounded in the behavioral foundations of innovation. I assess students' capacity for exploration through a multi-armed bandit task adapted from Ederer and Manso (2013), idea quality through an expert-rated scale calibrated against practicing innovators, and practical validation through a live shark tank-style competition. These outcomes reflect strategic search, analogical reasoning, and value-driven problem-solving—core cognitive behaviors emphasized in models of innovation. In addition, to examine mechanisms and distinguish innovation from other skills, I assess fluid intelligence, grit, emotional intelligence, and problem-solving using offline instruments suited to the field setting.

The intervention was implemented by Inqui-Lab Foundation, an NGO founded by engineers-turned-teachers who developed a structured pedagogy for underserved students, adapted from engineering design programs. The curriculum delivered 60 hours of training through weekly 90-minute sessions over two academic years. Students progressed through four blocks: analyzing existing technologies (e.g., analyzing the design evolution of everyday objects), identifying problems in their communities (e.g., school walkabouts to identify sustainability or local community challenges), designing and testing prototypes, and refining solutions based on feedback. Early sessions included critique-based discussions (e.g., comparing water filters or energy-efficient stoves), followed by guided problem-finding through school walkabouts and structured interviews. Final sessions focused on building prototypes using basic circuits and sensors. Two trained student leaders supported each classroom, and all schools received standardized innovation kits. Implementation fidelity was high: 98.1% of schools completed all sessions, and 88.0% of observed sessions delivered the curriculum in full. Control schools received only information about an annual innovation challenge with detailed criteria but no structured training, similar to statewide challenges previously conducted by the Telangana government (Government of Telangana, 2019; NITI Aayog, 2021). This created a research design that tests whether structured training improves innovation outcomes compared to self-directed learning.

The results from this evaluation reveal three key findings that collectively demonstrate the malleability of innovation-related skills and their relationship to traditional academic outcomes.

First, the intervention significantly improved students' innovation capabilities across multiple dimensions. Treated students demonstrated greater strategic exploration behavior in the multi-armed bandit task (0.096 SD, $p < 0.05$) and discovered higher-value

solutions (0.095 SD, $p < 0.05$) compared to control students. At the team level, treatment groups produced ideas rated significantly higher on quality (0.243 SD, $p < 0.01$) and perceived funding potential. These expert evaluations were validated through real-world outcomes: treatment teams were 24 percentage points more likely to secure funding (₹700 or $8) in an inter-school competition judged by external evaluators. These effects remain statistically significant after multiple-hypothesis testing corrections, providing robust evidence that innovation skills can be systematically developed.

Second, the program improved broader cognitive abilities that extend beyond domain-specific innovation skills. Students in treatment schools showed significant gains in fluid intelligence (0.120 SD, $p < 0.05$) and problem-solving capacity (0.122 SD, $p < 0.10$)—cognitive domains often considered difficult to influence through educational interventions. These improvements were particularly pronounced among students with higher baseline academic performance, suggesting that existing cognitive foundations may facilitate the development of these higher-order skills. Construct validity checks confirm that these cognitive measures capture distinct abilities from the innovation-specific outcomes and are not simply reflecting familiarity with task formats or increased motivation.

Third, the intervention produced opposite effects on math by baseline ability: gains for high-performers (+0.159 SD, $p < 0.05$) but declines for struggling students (-0.136 SD, $p < 0.05$). I find no significant average effects on math (0.013 SD, $p > 0.10$) and science test scores (0.022 SD, $p > 0.10$) across the full sample. But for students with above-median baseline math scores, the innovation program had complementary effects, improving both innovation skills and math performance. In contrast, for students with below-median baseline performance, the program improved innovation and problem-solving skills (0.098-0.114 SD) but reduced math scores (-0.136 SD, $p < 0.05$).

These results show that key behaviors associated with early-stage innovation—such as exploration, analogical reasoning, and adaptive problem-solving—can be cultivated through structured education. While most economic models of innovation primarily attribute cross-country differences in idea production to structural factors like market scale, access to capital, or institutional quality (Acemoglu and Cao, 2012; Kerr and Nanda, 2014; Furman et al., 2002), they give limited attention to the role of education in developing the underlying skills that drive innovation. In most models, education enters as a level of attainment (e.g., schooling, R&D labor), not as a producer of specific innovation-enabling higher-order skills. These findings suggest that idea generation is not just a function of environment and endowments—it can be shaped by exposure, pedagogy, and design. If ideas production has increasing returns to economic growth, then understanding how education contributes to their formation is not peripheral—it is central.

This research makes three distinct contributions to the economics literature. First, I contribute to the literature on higher-order skill development by providing rigorous evidence on an intervention that targets multiple cognitive and behavioral skills relevant for innovation. Human capital research shows that higher-order skills yield increasing

returns as technology transforms labor markets (Deming, 2022; Acemoglu and Autor, 2011; Cunha and Heckman, 2007). Recent studies demonstrate that targeted programs can develop metacognitive abilities (Alan et al., 2019; Alan and Ertac, 2018; Yeager et al., 2019; Alan and Mumcu, 2024; Bettinger et al., 2018), negotiation and business skills (Chioda et al., 2021; Ashraf et al., 2020b), and problem-solving skills (Bando et al., 2019; Ashraf et al., 2020a). I extend this work by showing that such skills can be cultivated in tandem through public education, even in resource-constrained environments.

Second, I contribute methodologically by addressing a central measurement challenge in the study of higher-order skills in developing countries. While economic theory recognizes non-cognitive and behavioral skills as key to productivity and labor market adaptability (Deming, 2022; Cunha and Heckman, 2008), credible tools for measuring these capacities in schools remain scarce. My study develops and validates a low-cost, scalable assessment strategy that draws on both economics and cognitive science, expanding the scope of what can be measured and for whom.

Third, I extend the innovation literature by examining who develops innovation skills and what problems they address. Research shows childhood exposure (Bell et al., 2019), access to knowledge (Biasi and Ma, 2022; Toivanen and Väänänen, 2016), and family background (Aghion et al., 2017) shape who becomes an inventor, creating substantial underrepresentation of disadvantaged groups. This matters because inventor demographics influence which problems receive attention—women inventors focus more on women's health (Koning et al., 2021) while inventors generally address problems familiar from their socioeconomic experiences (Einiö et al., 2023). The resulting innovation gaps disproportionately affect developing countries, where lost innovative potential may reduce output by up to 40% (Jones, 2022b). My study demonstrates that marginalized students not only develop innovation skills but create solutions specifically addressing challenges in their local contexts, providing evidence that diversifying innovation participants changes which problems receive attention—even at the educational stage.

Beyond the economics literature, this paper also speaks to policy tensions in India's education system. While the National Education Policy 2020 and Atal Tinkering Labs promote higher-order skills (India's Ministry of Education, 2020; NITI Aayog, 2021), classrooms remain shaped by exam preparation and memorization. As Muralidharan (2024) argues, this emphasis reflects schooling's function as a sorting mechanism rather than a vehicle for skill development. This divide between academic and applied skills is consistent with findings from Banerjee et al. (2025), who show that students who perform poorly on symbolic math tests can nonetheless solve equivalent problems in real-world contexts.

I find that high-performing students were able to transfer skills across domains, while lower-performing students reallocated effort toward hands-on tasks. Time-use data (Table 10) show that treated students did not reduce time spent on core academic activities like homework or subject periods, but instead reallocated time from unstructured

or "other" activities toward innovation and, for some, toward sports. The observed math declines for lower-performing students are therefore unlikely to reflect time crowd-out. Rather, these patterns point to a reallocation of cognitive effort within a fixed time budget—consistent with models of skill investment under comparative advantage (Becker, 1965). When presented with alternative learning opportunities, students who struggled with traditional academics redirected effort toward domains where they experienced greater success or engagement. These findings suggest that pedagogical design and perceived returns shape how students allocate attention, and that even academically lower-performing students can develop valuable skills when offered opportunities that align with their comparative strengths.

The remainder of the paper is organized as follows. Section 2 presents the theoretical framework, distinguishing between skill-based and idea-based education, and develops a growth model that motivates the intervention. Section 3 summarizes the context in which the program was implemented. Section 4 details the intervention design. Section 5 outlines the evaluation design. Section 6 gives a detailed account of the toolkit I use to evaluate the program. Section 7 describes the data. Section 8 presents my main results. In Section 9, I discuss the mechanisms through which the program might have improved higher-order skills and caused divergence in math achievement for the treatment group. Section 10 discusses cost-effectiveness and potential scale-up issues, and I conclude in Section 11.

## 2. Conceptual Framework: Education for Innovation vs. Basic Skills

In this section, I establish the theoretical distinction between traditional skill-based education and idea-based education to frame my investigation of how teaching innovation skills can drive economic growth. I define innovation broadly to include not just invention, but also adaptation and local problem-solving.

Classical endogenous growth theory (Romer, 1990) identifies the accumulation of new ideas ($A$) as the primary driver of sustained economic growth, rather than simply the expansion of labor or capital inputs. Romer's production function can be expressed as $Y = F(A,L,K) = A^{\sigma} \cdot L^{\alpha} \cdot K^{1-\alpha}$, where the stock of ideas $A$ enters with increasing returns ($\sigma > 0$) while labor exhibits diminishing returns ($\alpha < 1$). This fundamental difference—ideas generate increasing returns while labor does not—explains why knowledge accumulation drives long-run growth.Ideas are "non-rival" - once created, they can be used by many people simultaneously without being depleted, unlike "rival" inputs like labor that can only be used in one place at one time. This non-rival property means a society's income per person grows with the total stock of knowledge in the economy, not merely with ideas possessed by each individual.

This distinction in growth theory maps directly onto different educational approaches. Skill-based education primarily enhances labor productivity ($L$) by training students to

Table 1: Contrasting Skill-Based and Idea-Based Education

| Type of Education | Cognitive Focus | Output | Economic Role | Example |
|---|---|---|---|---|
| **Skill-based** | Procedural, routine, task-specific | Known solutions | Raises $L$ — labor productivity with diminishing returns | Learning how to *assemble* a walking stick using a manual |
| **Idea-based** | Generative, creative, exploratory | Novel solutions | Raises $A$ — knowledge creation with increasing returns | Imagining a *smart* walking stick with GPS and reminders |

perform known tasks efficiently—to follow instructions, replicate techniques, or operate tools. In contrast, idea-based education cultivates the capacity to generate new knowledge ($A$) through exploration, analogical reasoning, problem identification, and divergent thinking. Table 1 summarizes these contrasting approaches. While traditional educational policy in low- and middle-income countries emphasizes basic skills such as literacy and numeracy, this framing positions education as a mechanism for *catch-up growth* through improved labor efficiency, not as a catalyst for *frontier-expanding* innovation.

I extend the standard Romer-style knowledge accumulation function to incorporate the role of idea-based education, modeling how investments in $E_I$—education focused on creativity, exploration, and reasoning—increase the productivity of both innovation and diffusion activities:

$$\dot{A} = \nu \cdot (1 + \delta_A E_I) \cdot L_A \cdot A^{\phi} + \mu \cdot (1 + \delta_D E_I) \cdot L_D \cdot A$$

In this extension, $L_A$ represents labor allocated to idea generation, $L_D$ represents labor involved in adapting existing knowledge, and $\delta_A$ and $\delta_D$ capture the effect of idea-based education on each channel. This formulation allows $E_I$ to influence the growth trajectory even before individuals fully master foundational skills, enhancing knowledge production across the economy—not just within formal R&D settings, but in everyday problem-solving contexts.

A key implication of this framework is that meaningful idea generation does not necessarily require mastery of the traditional curriculum. My field data reveals that student teams exposed to idea-based education have produced conceptually rich innovations without advanced formal knowledge. For example, one student team designed an "Elderly Walking Stick" (Figure A2)—a cane embedded with a medication box, GPS tracking, and an audio alert system for elderly individuals. This innovation emerged not from mastery of engineering or biomedical science, but from the team's ability to observe, empathize, and recombine familiar objects into a novel solution addressing a real community need. See similar examples in Appendix A

Traditional educational systems focus overwhelmingly on transmitting established

knowledge and developing procedural skills, with a common assumption that foundational skills must precede any higher-order thinking or creativity. This paper examines whether this sequencing is necessary—specifically, whether innovation skills can be effectively taught, whether they must follow mastery of basic skills, or whether they can develop in parallel with foundational learning. Moreover, this theoretical framework suggests that teaching idea generation may yield increasing returns to economic growth, while teaching basic skills alone faces diminishing returns—making this question particularly consequential for educational policy.

The remainder of the paper empirically tests these assumptions—particularly whether innovation skills can be developed in parallel with foundational learning, and whether doing so yields measurable returns.

## 3. Context, Intervention, and Delivery

### 3.1 Context

India faces a fundamental human capital allocation challenge: despite rapid growth in high-skill sectors, most workers remain in low-productivity traditional industries. Only 7% of young people complete higher education (AISHE, 2020), while 88% of workers enter small enterprises (ILO, 2018).

This situation stems from three interconnected crises. India confronts a severe learning crisis in basic education, with only 50% of fifth-grade students able to read at a second-grade level (ASER, 2022). The country simultaneously faces persistent youth unemployment, reaching 23.2% among those aged 15-29 (PLFS, 2022). Even those who complete technical education face the third crisis: an employability gap where 80% of engineering graduates require significant additional training before being "employable" (India Skills Report, 2022), reflecting a disconnect between learning and workforce needs.

India's economic goals are ambitious but human capital development is not keeping pace with these economic challenges. Initiatives like Make in India, Startup India, and the National Artificial Intelligence Strategy assume the availability of innovation-capable workers across sectors. However, the traditional education system—focused primarily on rote learning and standardized testing—struggles to develop higher-order thinking skills these economic initiatives require.

To address this gap, several innovation-focused educational initiatives have emerged. The Atal Innovation Mission has established over 10,000 Atal Tinkering Labs in schools nationwide with another 10,000 in process, representing a $240 million investment aimed at creating "one million young innovators" (NITI Aayog, 2021). Delhi's Entrepreneurship Mindset Curriculum ($27 million annually) targets similar skills at scale. Despite these investments, little causal evidence exists on their effectiveness.

Telangana exemplifies India's economic duality: Hyderabad's tech sector generates 30% of state GDP with just 7% of workers, while small enterprises (38% of workers, 23%

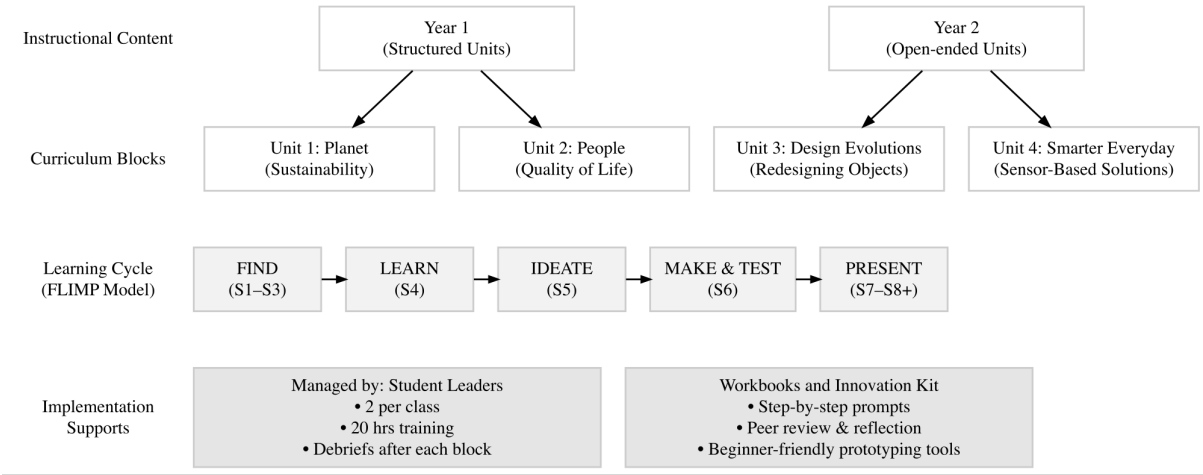



Figure 1: Structure of the Innovation Training Program

of GDP) and agriculture (55% of workers, 18% of GDP) operate at far lower productivity levels (Government of Telangana, 2022). The state has responded with its 2015-2025 Innovation Policy, an annual School Innovation Challenge engaging 3,000+ schools, and infrastructure like T-Works, the state-sponsored prototyping center and makerspace supporting hardware innovation and skills development (Government of Telangana, 2019).

These innovation education efforts could particularly benefit marginalized communities by developing applied thinking skills relevant across economic sectors. The residential schools in this study serve exclusively Scheduled Castes and Scheduled Tribes—constitutionally protected groups comprising 24.5% of Telangana's population but representing less than 5% of technology sector employment. These communities face persistent barriers to economic mobility (Asher et al., 2022; Hnatkovska et al., 2013) and may benefit from skills that enhance productivity in traditional sectors while potentially opening pathways to higher-value economic activities.

This research examines a structured innovation education program at a critical juncture—8th and 9th grades—when Indian students begin specializing before high-stakes examinations that determine their economic trajectories. The study provides initial evidence on whether such programs can develop economically valuable skills among students historically excluded from India's high-productivity sectors.

### 3.2 Intervention and Delivery

The innovation training program followed a two-year structure, reaching 2,500 Grade 8 students across 40 residential government schools in Telangana. Students participated in

40 weekly sessions of 90 minutes each, totaling 60 instructional hours over two academic years.

The curriculum was explicitly structured around a five-step innovation cycle: FIND (identifying local problems), LEARN (researching existing solutions), IDEATE (brainstorming solutions), MAKE & TEST (prototyping), and PRESENT (sharing ideas and integrating feedback). This approach aligns with Razzouk and Shute (2012)'s conceptualization of design thinking as "an analytic and creative process that engages a person in opportunities to experiment, create and prototype models, gather feedback, and redesign." Each of the four curriculum units—*Planet*, *People*, *Design Evolutions*, and *Smarter Everyday*—was grounded in this cycle and spanned 8–10 sessions. Students applied these steps through activities such as conducting school resource audits (FIND), reviewing real-world case studies (LEARN), generating and evaluating ideas using rubrics (IDEATE), building prototypes using provided kits (MAKE & TEST), and presenting to peers or school communities (PRESENT).

Year 1 emphasized building foundational skills. Students developed a keen eye for identifying community challenges, learned to investigate causes through structured tools, and gained confidence using a step-by-step problem-solving approach. In Year 2, students were expected to demonstrate increased autonomy, take ownership of the innovation process, and design solutions using sensor technologies for real-world problems. Unit themes were intentionally sequenced to support this progression: while Year 1 units focused on sustainability and quality of life, Year 2 units centered on product redesign and technology integration.

Although the five-step cycle remained constant, the openness of each task varied across years. In Year 1, students worked within guided thematic contexts. For example, in the *Planet* unit, they identified local sustainability issues, mapped school water usage, and conducted waste audits. In *People*, they conducted interviews and developed empathy-driven solutions to improve daily life in their communities. In Year 2, students took on more open-ended tasks. In *Design Evolutions*, they analyzed how objects evolve to meet user needs and redesigned items such as chalk dusters or school furniture. In *Smarter Everyday*, they integrated light, sound, and motion sensors into functional prototypes such as safety devices, irrigation controllers, and wearable monitors. Final sessions of each unit allowed students to independently define problems, plan prototypes, and iterate based on user testing and peer feedback.

Sessions were delivered through structured student workbooks containing step-by-step prompts, reflection activities, and peer feedback mechanisms. These workbooks enabled students to collaborate and make progress semi-autonomously, even without formal instructors. Each classroom also received an "innovation kit" valued at Rs. 15,000 ($180 USD), containing circuits, sensors, recyclable materials, and beginner-friendly prototyping tools aligned to the curriculum's technical demands.

Two student leaders nominated by teachers facilitated each classroom after receiv-

ing 20 hours of training. These leaders supported material handling, timekeeping, and team collaboration, and participated in debriefing sessions after each curriculum block to standardize implementation. Leaders were instructed to answer only clarifying questions, preserving students' ownership of the learning process.

Implementation fidelity was monitored through weekly observations by trained field staff team of seven 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 schools, 98.1% completed every required session, and 88.0% of observed sessions implemented all core components. These high adherence rates strengthen the program's internal validity and suggest feasibility for scale. Details on fidelity protocols, sample workbooks, and session designs appear in Appendix A.1.

## 4. Evaluation Design and Timeline of the Study

I selected 80 schools from a pool of 428 eligible schools based on operational feasibility for the implementation partner and input from the state government ensuring no other related programs were occurring in these schools during the study period. Within this selected group, I employed a stratified randomization design. Stratification was based on three key dimensions: geographical setting (urban/rural), gender composition (girls'/boys' schools), and administrative affiliation (social welfare/tribal welfare departments). I randomly assigned 40 schools to the treatment group and 40 to the comparison group within each stratum.[2]

The intervention consisted of structured innovation training provided to the treatment group on a weekly basis. The comparison group did not receive this training. Instead, at the beginning of each academic year, comparison schools were formally notified about an upcoming 'Shark Tank Challenge' via an official letter. This communication emphasized that teams would be evaluated based on how beneficial their innovative idea was for its intended users. The letter also provided basic competition logistics but offered no guidance on idea development or preparation. This innovation challenge concept was familiar to students due to previous state-organized innovation competitions (Telangana State Innovation Cell, 2021; AMS Public Relations Society, 2022) and the widespread popularity of entrepreneurship reality shows in India (Kushagra, 2023; Soni and Sharma, 2022; Wadhwa, 2022).

As shown in Figure 2, both treatment and comparison schools participated in the same competition timeline from August 2022 to November 2024, with baseline, midline,

[2]An a priori power analysis was conducted for this cluster-randomized trial. With 80 schools and an assumed average of 50 students per school, I calculated 80% power to detect an effect size of approximately 0.15 standard deviations at the 5% significance level (two-tailed), assuming an intraclass correlation coefficient of 0.15 and that baseline covariates would explain about 30% of the outcome variance. Full details of the power calculation are provided in Appendix B.6.

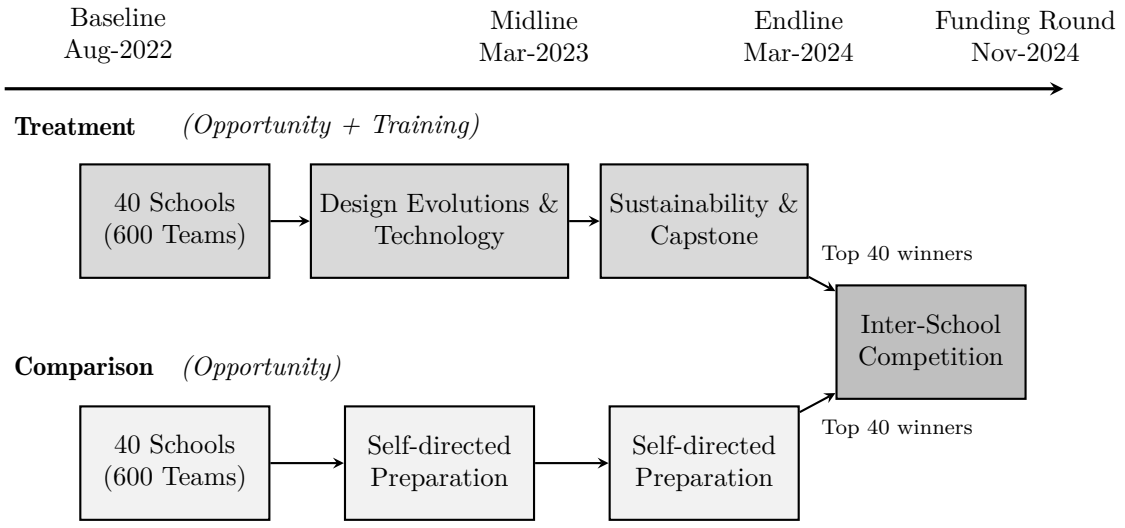



FIGURE 2: PROGRAM TIMELINE

*Notes:* The timeline illustrates that 80 schools were divided into treatment and comparison groups using strata randomization. In every school, students were split into fifteen teams, with each team consisting of 4 students chosen at random. This procedure is repeated in comparison schools as well. Treatment schools received structured design thinking curriculum delivered through workbooks and peer facilitators. Comparison schools had the same opportunity to develop ideas for the competition but without specialized training. Both groups consist of similar residential government schools.

and endline assessments. The key difference was in their preparation approach. While treatment schools received 40 structured sessions totaling 60 hours of formal training over two years, comparison schools engaged in "Self-directed Preparation" alongside the standard school curriculum. Since these were residential schools, students in both groups had independent homework periods after regular school hours. Treatment schools scheduled their weekly 90-minute sessions during independent study periods, ensuring no academic subject time was displaced. Comparison group teams were encouraged to use this same time to develop their innovation ideas on their own. Thus, both groups utilized the same independent study time allocation, but the treatment group received structured training while the comparison group engaged in self-directed preparation. Ultimately, top performers from both groups would compete in the annual Inter-School Competitions, which were organized in October 2023 and November 2024.

Both groups were thus motivated by the prospect of a competition, but only the treatment group received structured training. This design allows me to assess whether formal innovation training leads to better outcomes compared to students' self-directed efforts in preparing for a familiar type of innovation challenge.

The experiment spans two academic years, with data collection at three primary points: baseline in August 2022, midline in March 2023 (end of Year 1), and endline in March 2024 (end of Year 2). An inter-school competition for winning teams from each school was conducted in November 2024, in the following academic year, due to logistical constraints and school calendar considerations. The map in Figure 6 shows the geographic distribution of treatment and comparison schools across 12 districts in Telangana state.

## 5. Outcomes

The primary outcome of interest is innovative ability—defined as students' capacity for strategic exploration, idea generation, and applied problem-solving. I measure this construct using three complementary approaches: a behavioral task (multi-armed bandit), an expert-rated idea quality scale, and a real-world pitch competition. I also measure a set of related higher-order skills—fluid intelligence, grit, emotional intelligence, and structured problem-solving—to explore mechanisms and distinguish innovation from correlated cognitive traits.[3] Finally, I assess math and science performance to examine potential spillovers to academic domains, even though achievement was not a direct target of the program.

Recognizing the challenge of measuring innovative abilities in young students, I employ three distinct measures at individual and team levels. First, I use a lab-in-the-field game based on March (1991)'s seminal work on exploration and exploitation in innovation processes, using a multi-arm bandit problem to assess individual innovative behavior. Second, all students collaborate in teams to submit ideas, which are evaluated using an idea quality measure developed with a survey of experienced innovators. This measure has established predictive validity for real-world innovation outcomes such as user reach and funding raised. Third, the top-performing teams, selected based on their idea quality, participate in an inter-school funding competition. This competition features expert jury assessment, including non-profit entrepreneurs, social impact investors, and corporate innovators, employing a double-blind process with two jurors evaluating every 20 teams, balanced across treatment and control groups.

### 5.1 Innovative Ability Using Multi-Armed Bandit Problem

This study uses a multi-armed bandit task to measure innovative ability, operationalized as students' capacity to explore new ideas, learn from feedback, and ultimately discover valuable outcomes. I ground this approach in March (ibid.)'s exploration-exploitation framework, which has informed experimental paradigms in economics, cognitive science, and organizational behavior (Ederer and Manso, 2013; Herz et al., 2014; Bandiera et al., 2021). I designed the task around business strategy rather than product innovation to ensure that innovative behavior could be observed independently of the treatment content, avoiding overlap with product-based themes and enabling a cleaner behavioral measure of exploration.

The task simulates a business strategy challenge: over 12 rounds, students choose combinations of stand type, location, and other business parameters to maximize hourly profit. Each decision yields deterministic feedback, and students view their cumulative

[3]There are various terms used in the literature to describe non-academic abilities, including 'non-cognitive skills,' 'soft skills,' and 'socio-emotional skills.' Following Deming (2022), I use the term 'higher-order skills' to emphasize that these abilities involve complex cognitive processes that extend beyond basic academic knowledge. This terminology also avoids the misleading cognitive/non-cognitive dichotomy since higher-order skills often involve both cognitive and socio-emotional components

results on-screen. While the full configuration space is finite and observable, students are not told the underlying profit function and must learn through experience. Each decision yields probabilistic feedback from a fixed distribution tied to the selected stand type. Students are not told the profit function but learn through repeated feedback, enabling analysis of strategic behavior in a known but opaque environment.

The task includes three business types with distinct risk-reward profiles. The Fresh Juice Stand represents a safe exploitation strategy, with a high expected value (EV = ₹43.27) and lower variance. In contrast, Ice Cream and Pani Puri Stands represent high-risk, high-reward options—with substantially higher standard deviations and maximum possible profits (up to ₹250). Students are incentivized to identify the single most profitable configuration in their class, encouraging goal-directed exploration under uncertainty.

To quantify exploratory behavior, I construct a composite Exploration Index from two components: whether a student chose a different stand type than the example (Prasuna's Fresh Juice Stand) in Round 1, and how many distinct combinations of stand type and location they tested across rounds. The first captures early divergence from exemplars; the second reflects breadth of conceptual search. I standardize and average these components into a single continuous index.

To assess performance, I use the maximum profit achieved by each student across the 12 rounds. This metric reflects whether students identified high-value configurations through search and adaptation. Given the payoff structure, maximum profit is a direct, interpretable measure of realized success.

In addition to these continuous measures, I also define a binary profile for Systematic Innovators based on three behaviorally grounded criteria. First, the student must have selected a high-risk stand (Ice Cream or Pani Puri) in at least half of the 12 rounds. Second, after receiving a low-profit outcome—defined relative to their own prior best—the student must have returned to the same risky stand type in at least one later round. This condition captures strategic persistence in a specific risk space following underperformance. Third, the student must have achieved a maximum profit greater than ₹30, exceeding the benchmark shown in the task example. This profile reflects not just broad exploration or outcome success alone, but sustained engagement with risk, adaptation after negative feedback, and eventual high-value discovery.This profile captures those who explore broadly and succeed—providing a behavioral proxy for strategic innovation.

Unlike stochastic bandit tasks, this implementation features deterministic feedback and a fixed option space. This design allows me to observe meaningful variation in strategic behavior—such as deviation from the safe default, persistent engagement with high-risk arms, and continued search even after finding profitable configurations. The outcomes thus reflect not only what students chose, but how they navigated the trade-offs central to innovation.

The Appendix (Section B.1) provides further details on the task design, including the

exact profit distributions used for each stand type, an example decision sequence, and screenshots of the task interface. I also describe the pilot testing process and how task elements such as input persistence and profit feedback were implemented. To encourage effort, students were informed that the highest scorer in each class would receive a prize.

### 5.2 Innovative Ability Using Quality of Ideas

Building on the multi-armed bandit game, I developed a complementary measure to address the limitations in capturing real-world innovation ability. I created the Innovation-S scale, a novel psychometric tool designed to measure the quality of innovative ideas.

The Innovation-S scale was developed through a process involving in-depth interviews with 14 successful innovators, including founders, scientists, and a Nobel Laureate, followed by a survey of 44 experienced innovators from diverse fields. The participants represented a range of industries, with a significant concentration in education (40.9%) and technology (15.9%), and included founders (25%), product managers (13.6%), and analysts (11.4%) among others. This diverse group had experience across multiple innovation stages - 66% in ideation, 61% in development, 55% in launching, and 59% in improvement, providing a balanced perspective across different organizational structures.

The resulting scale has a 2-factor structure (novelty and value-add) comprising seven items, such as "offers a new solution to a problem" and "improves the lives of those who will use it." The scale demonstrates valid psychometric properties (Cronbach's alpha = 0.69) and strong predictive validity, with correlations of 0.71 and 0.31 with user base and funding respectively. These metrics indicate that the Innovation-S scale is a reliable tool for measuring innovative ability, capable of predicting real-world innovation outcomes across various scales of operation and funding levels.

To apply this scale in evaluating student ideas, I carefully selected and trained two expert evaluators from a pool of over 1,200 applicants. These evaluators independently rated each idea submitted by both treatment and comparison teams using the Innovation-S scale and provided a "willingness to fund" measure for each idea. Appendix B.2 describes the full development and validation process of the Innovation-S scale, including factor structure, item selection, and external validation results. Appendix B.3 describes how the team idea was collected, processed, and evaluated to ensure consistency across teams.

### 5.3 Innovative Ability Using Interschool Fundraising Challenge

To assess the practical impact of the innovation training program, I implement a real-world fundraising challenge as the third outcome measure. From the pool of ideas collected at the end of the year, I selected the top 40 ideas from each group based on the raters' ratings, resulting in 80 total ideas for the challenge. To minimize preparation bias and simulate real-world constraints, participants receive only 5 days' notice before the challenge.

The evaluation process involves an expert jury comprising eight individuals: non-profit

entrepreneurs, social impact investors, and corporate innovators. I employ a double-blind assessment method, where each pair of jurors evaluates 20 teams, balanced across treatment and control groups. The key outcome measure is the difference in average fundraising amounts between the treatment and control groups. Appendix B.4 provides additional detail on jury composition, scoring rubrics, and procedures to ensure double-blind evaluation and real-stakes funding decisions.

### 5.4 Related Higher Order Skills

To examine mechanisms and distinguish innovation from correlated skills, I measure four higher-order constructs: fluid intelligence, problem-solving, emotional intelligence, and grit. Each was assessed using adapted versions of validated instruments designed for use in school settings. Further details on adaptation, piloting, and cultural validation are provided in Appendix B.5.

**Problem-Solving.** The intervention required students to identify problems, apply constraints, and test structured solutions—processes that parallel formal problem-solving in well-defined domains. Problem-solving refers to the ability to operate within established parameters to reach optimal solutions, typically with known goals and methods (Newell and Simon, 1972; Russell and Norvig, 2010). I assess this skill using timed 4×4 Sudoku puzzles, which preserve the structure of NP-complete problems while being developmentally appropriate (Yato and Seta, 2003). Measuring problem-solving helps distinguish between structured reasoning and exploratory innovation, which may develop along different cognitive paths (March, 1991; Simon, 1973).

**Fluid Intelligence.** The program repeatedly engaged students in abstraction, analogical reasoning, and unfamiliar problem-solving contexts—cognitive demands that may plausibly activate fluid reasoning during adolescence, a period of heightened neurocognitive plasticity (Jauk et al., 2013). Fluid intelligence refers to the ability to reason, identify patterns, and solve novel problems independent of acquired knowledge (Carpenter et al., 1990). I measure this skill using Raven's Progressive Matrices (Raven, 2000), a widely used non-verbal test designed to minimize the influence of cultural or linguistic background. Measuring fluid intelligence allows me to test whether the program affected core cognitive capacities, and to examine whether these changes help explain improvements in innovation-related outcomes.

**Emotional Intelligence.** The program's emphasis on teamwork, peer feedback, and structured presentations offered students repeated opportunities to engage in collaborative work—potentially developing emotion recognition and social cognition. Emotional intelligence refers to the ability to perceive, understand, and interpret emotional states in others, which can support group-based innovation processes (Baron-Cohen et al., 2001; Vellante et al., 2013). I measure emotional intelligence using the Reading the Mind in

the Eyes Test (RMET), a widely used instrument for assessing empathic accuracy and social perception. This measure allows me to assess whether the collaborative structure of the program influenced interpersonal dimensions of higher-order skill development.

**Grit.** The long-term, iterative structure of the program—problem-finding, prototyping, and refinement—was designed to reward persistence and sustained effort despite failure. Grit is defined as perseverance and passion for long-term goals (Duckworth et al., 2007), and has been linked to both educational and innovation-related outcomes (Galton, 2019; Jachimowicz et al., 2018). I assess grit using a behavioral choice task adapted from Alan et al. (2019), where students choose between easy and difficult puzzles with differential rewards. This revealed-preference measure avoids the limitations of self-report and provides insight into how the program may have influenced persistence in the face of cognitive challenge.

Further details on adaptation, piloting, and cultural validation are provided in Appendix B.5.

### 5.5 Academic Outcomes and Motivation

Academic ability has been shown to predict innovation outcomes, particularly in high-skill and STEM fields (Bell et al., 2019). To assess spillover effects, I measure math and science achievement using standardized assessments adapted from school textbooks and the Mindspark study (Muralidharan et al., 2019). These assessments were validated on out-of-sample students and aligned with the Indian curriculum.

I also collect self-reported data on attendance, daily time use, and classroom engagement (e.g., questions asked to teachers) to capture broader behavioral impacts. These indicators reflect potential pathways through which innovation education might influence student engagement, social connection, and future orientation (Yeager et al., 2019; Hoxby, 2000; Zárate, 2022; Sacerdote, 2014). Although not primary outcomes, these measures help contextualize how students respond to a pedagogically distinct program.

‘

## 6. Data and Empirical Analysis

Based on official school registry data, the total potential sample included 5,871 registered students across the 80 schools, with approximately equal numbers in treatment and control schools. At baseline (August 2022), I collected data from 4,948 students, representing 84.9% of registered students in the control group and 83.7% of registered students in the treatment group. At midline (March 2023), my sample included 5,077 students, representing 87.9% of registered control students and 85.1% of registered treatment students. By endline (March 2024), I had data from 4,825 students, representing 79.9% of registered control students and 84.5% of registered treatment students.

### 6.1 Outcome Measures by Data Collection Wave

I collected a wide range of outcome measures across three data collection waves. At baseline (August 2022), I measured mathematics and science achievement, Ravens Progressive Matrices (fluid intelligence), Reading Mind in Eyes Test (emotional intelligence), Big Five personality traits, risk preferences, and innovation task performance (multi-armed bandit task). I also collected data on social networks (number of friends) and career aspirations.

At midline (March 2023), I collected the same set of academic achievement measures (math and science scores), cognitive assessments (Ravens and RMET), innovation task performance, and Big Five personality traits. I also gathered data on subject preferences across eight school subjects, risk preferences, attendance, classroom participation, and updated social network and career aspiration measures.

At endline (March 2024), I expanded the assessment battery to include problem-solving skills and grit in addition to the previously collected measures of academic achievement, fluid and emotional intelligence, and innovation task performance. I added detailed time allocation measures across various activities (school, homework, sports, innovation, other) and collected qualitative data including creative writing samples, innovation improvement suggestions, and evaluations of external innovation ideas. I also collected one innovation idea from each treatment and control team, resulting in 1,191 ideas for evaluation (564 from comparison and 627 from treatment teams).

To enable covariate adjustment and improve precision, I constructed a socioeconomic status (SES) index using principal component analysis based on eight household asset indicators.

All outcome measures at endline were administered digitally using platforms with mandatory input fields, resulting in 100% data completeness among assessed students across all primary and secondary outcomes.

### 6.2 Internal Validity

Table 2 presents baseline characteristics and balance checks across treatment and control groups. The first panel shows school-level characteristics and confirms exact balance on stratification variables (tribal school, girls' school, rural location). Other school-level differences are not statistically significant, and the joint test yields an F-statistic of 0.089 ($p = 0.999$), confirming balance at the cluster level.

Panel B reports student-level characteristics and baseline outcome measures. Balance is strong overall, with no significant differences on key academic or innovation measures. The socioeconomic status (SES) index shows a small but statistically significant difference (difference = 0.17, $p = 0.022$), which I adjust for in all outcome regressions. The joint test for covariate balance among students yields $F = 0.949$ ($p = 0.504$), indicating no systematic imbalance across groups.

Panel C examines sample composition and attendance across the three waves. While attendance varied slightly by group and wave, the percentage of students observed in all

Table 2: Balance Table

| Variable | All | Treatment | Control | Difference | p-value |
|---|---|---|---|---|---|
| **Panel A: School Characteristics** | | | | | |
| Tribal School | 50.00 | 50.00 | 50.00 | 0.00 | 1.00 |
| Girls School | 50.00 | 50.00 | 50.00 | 0.00 | 1.00 |
| Rural School | 62.50 | 62.50 | 62.50 | 0.00 | 1.00 |
| Library | 50.00 | 50.00 | 50.00 | 0.00 | 1.00 |
| Playground | 13.75 | 15.00 | 12.50 | 2.50 | 0.75 |
| Ramp Facility | 70.00 | 67.50 | 72.50 | -5.00 | 0.63 |
| Water Facility | 1.00 | 1.00 | 1.00 | 0.00 | . |
| Computer Lab | 66.25 | 65.00 | 67.50 | -2.50 | 0.82 |
| Toilets | 61.26 | 58.12 | 64.40 | -6.28 | 0.60 |
| Observations | 80 | 40 | 40 | | |
| F-statistic: 0.089 | | p-value: 0.999 | | | |
| **Panel B: Student Characteristics and Baseline Outcomes** | | | | | |
| Age | 15.48 | 15.46 | 15.49 | -0.03 | 0.643 |
| SES Index | 2.33 | 2.42 | 2.25 | 0.17 | 0.022 |
| Mother's Occupation (Housewife) | 0.39 | 0.40 | 0.37 | 0.03 | 0.426 |
| Father's Occupation (Farmer) | 0.68 | 0.65 | 0.71 | -0.06 | 0.167 |
| Computer Skills | 34.33 | 35.13 | 33.54 | 1.59 | 0.681 |
| Number of Friends | 4.77 | 4.80 | 4.74 | 0.06 | 0.756 |
| Math Score | 41.72 | 41.24 | 42.21 | -0.97 | 0.637 |
| Science Score | 53.63 | 53.61 | 53.64 | -0.03 | 0.984 |
| Fluid Intelligence | 58.49 | 58.31 | 58.67 | -0.36 | 0.833 |
| Emotional Intelligence | 44.32 | 43.96 | 44.68 | -0.72 | 0.451 |
| Innovation Skills | 45.66 | 45.97 | 45.35 | 0.62 | 0.683 |
| Risk Preference Index | 0.00 | -0.02 | 0.02 | -0.04 | 0.639 |
| Observations | 5863 | 2935 | 2928 | | |
| F-statistic: 0.949 | | p-value: 0.504 | | | |
| **Panel C: Sample Composition and Attrition** | | | | | |
| Present at Baseline (%) | 81.63 | 79.80 | 83.47 | -3.67 | 0.072 |
| Present at Midline (%) | 84.85 | 84.05 | 85.66 | -1.61 | 0.270 |
| Present at Endline (%) | 77.64 | 79.15 | 76.13 | 3.02 | 0.235 |
| Present in All Waves (%) | 59.15 | 58.47 | 59.84 | -1.37 | 0.637 |

*Notes:* This table presents balance tests for school characteristics, student characteristics, and baseline outcome variables. Panel C shows sample composition and attrition patterns across waves. The last two columns show the difference between treatment and control means and the p-value from a t-test of equality of means. The F-statistic and corresponding p-value are from a joint test of significance of all variables in predicting treatment assignment within each panel. All academic and cognitive measures are standardized to have mean 50 and standard deviation 10 in the control group.

three waves is similar across treatment and control groups (58.5% vs. 59.8%, $p = 0.637$). To address any concern about selective attrition, I report baseline balance restricted to the sample of endline attendees in Table 11. Balance remains strong in this subsample, and the joint F-test again fails to reject equality ($p = 0.408$). Furthermore, tests for differential attrition—interactions between treatment and baseline characteristics—yield a joint $\chi^2(3) = 0.25$ ($p = 0.969$), indicating that the treatment did not systematically affect who was observed at endline.

Together, these findings support the internal validity of my experimental design. Baseline equivalence holds both in the full sample and among endline respondents, and there is no evidence that treatment status predicts attrition or compositional change in the es-

timation sample. To preserve sample size and covariate balance, I use mean imputation for missing baseline covariates among students observed at endline (136 students).

### 6.3 Empirical Strategy

To estimate the impact of the innovation training program, I use the following regression specification for individual-level outcomes:

$$Y_{is} = \alpha + \beta \text{ Treatment}_s + \gamma \text{ Strata}_s + \delta Y_{is}^{baseline} + \phi X_{is} + \psi S_s + \varepsilon_{is} \quad (1)$$

where $Y_{is}$ represents the outcome of interest for student $i$ in school $s$, $\text{Treatment}_s$ is an indicator for whether school $s$ was assigned to the treatment group, $\text{Strata}_s$ is a vector of stratification variables (indicators for tribal schools, girls' schools, and rural schools), $Y_{is}^{baseline}$ represents baseline controls for the same outcome when available, $X_{is}$ is a vector of individual-level controls (age, socioeconomic status, computer skills, and parent occupation categories), and $S_s$ represents school-level controls (building status, toilet facilities, wall, library, playground, water access, medical facilities, and accessibility ramps). Standard errors are clustered at the school level to account for within-school correlation in outcomes.

For outcomes that were not collected at baseline (problem-solving skills and grit), I omit the baseline control but include the full set of other controls and stratification variables.

For team-level outcomes, particularly the quality of innovation ideas collected at endline, I use a similar specification:

$$Y_{ts} = \alpha + \beta \text{ Treatment}_s + \gamma \text{ Girls}_s + \lambda \text{ Tribal}_s + \psi S_s + \phi X_{ts} + \varepsilon_{ts} \quad (2)$$

where $Y_{ts}$ represents the outcome for team $t$ in school $s$, and other variables are defined as above. The vector $X_{ts}$ includes team-level characteristics such as team size and average socioeconomic status, constructed by averaging the individual-level covariates of team members.

Given the multiple outcomes examined in this study, I address concerns about false positives by implementing two approaches to multiple hypothesis testing. First, I use the Romano-Wolf step-down procedure (Romano and Wolf, 2005) to adjust p-values across families of outcomes. Second, I implement the wild cluster bootstrap procedure (Cameron et al., 2008) to obtain more accurate p-values for my primary treatment effect estimates, given the cluster-randomized design. I also explore heterogeneity in treatment effects by gender, ses, and baseline academic performance.

### 6.4 Differential Attrition and Attendance as an Outcome

I work in a setting where student absenteeism is common due to various factors including family obligations, seasonal work, and health issues. The sample comprises 5,863 students across 80 schools. Attendance patterns varied across waves, with control schools showing

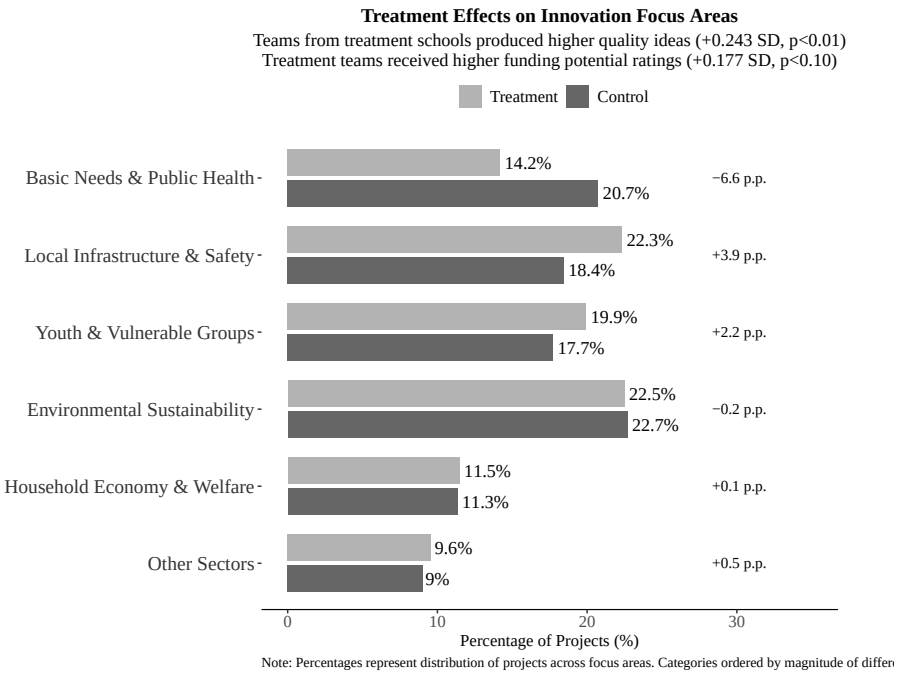



FIGURE 3: TREATMENT EFFECTS ON INNOVATION FOCUS AREAS

higher participation at baseline (83.5% vs. 79.8%), similar rates at midline, and lower attendance at endline (76.1% vs. 79.1%).

This differential attendance could be interpreted in two ways. First, as a positive outcome of the intervention - the innovation training appears to have increased student engagement with school, leading to improved attendance. Second, as a potential threat to the internal validity of my treatment effect estimates on other outcomes, if the composition of students present at endline differs systematically between treatment and control schools.

To address this second concern, I conducted an extensive analysis of baseline characteristics among students present at endline. Among endline attendees, treatment and control groups remain well-balanced on all key baseline measures as shown in Table 11. The joint F-test confirms no systematic differences between treatment and control students in the analyzed sample. I also examined whether baseline characteristics predict differential attrition by treatment status through interaction tests. The joint test of interactions between treatment and baseline academic and innovation measures yields a chi-squared of 0.25 ($p = 0.969$), demonstrating that the intervention did not fundamentally alter the relationship between student characteristics and attendance patterns. This suggests that while attendance patterns differed between treatment and control schools, the composition of students remained comparable.

Table 3: Treatment Effects on Innovation Outcomes (Multi-Armed Bandit Task)

| | Treatment Effect (1) | Wild p-val (2) | MHT p-val (3) | Within p-val (4) | Obs. (5) |
|---|---|---|---|---|---|
| Exploration Index | 0.093*** | 0.002 | 0.000 | 0.000 | 4531 |
| | (0.026) | | | | |
| Maximum Value Discovery | 0.092** | 0.024 | 0.000 | 0.000 | 4531 |
| | (0.038) | | | | |
| Systematic Innovator | 0.043** | 0.034 | 0.000 | 0.000 | 4531 |
| | (0.018) | | | | |

*Notes:* This table reports treatment effects from OLS regressions of the specified outcomes on treatment status. All regressions control for baseline measures, strata fixed effects, and demographic characteristics. Standard errors clustered at the school level are reported in parentheses. Exploration Index is a composite measure of exploratory behavior. Maximum Value Discovery is the highest profit found. Systematic Innovator = high in both.

## 7. Results

This section presents the estimated treatment effects of the two-year innovation education program on students' outcomes. I organize the presentation by outcome category: innovation skills and higher-order cognitive skills, academic performance, and student behavior.

### 7.1 Treatment Effects on Innovation Skills

Table 3 presents treatment effects on individual innovation measures using the multi-armed bandit task. Treatment students scored 0.10 SD higher on the Exploration Index ($p = 0.002$), demonstrating stronger tendencies to explore multiple options early in the decision process. They also achieved higher performance, scoring 0.10 SD higher on Maximum Value Discovery ($p = 0.022$), indicating greater ability to identify high-reward configurations through learning and adaptation. Both effects remain statistically significant after adjusting for multiple hypothesis testing and using wild cluster bootstrap $p$-values.

Additionally, treatment students were 4.7 percentage points more likely to be classified as Systematic Innovators ($p = 0.001$)—those who both explored widely and successfully identified high-value strategies. These findings suggest that the program improved both the cognitive strategy and execution of innovation behavior in uncertain environments.

Table 4 presents treatment effects on team innovation outcomes, with all variables standardized relative to the control group. Panel A shows that teams from treatment schools scored 0.26 SD higher on the overall Innovation-S scale ($p = 0.000$), with significant improvements on both the novelty dimension (0.18 SD, $p = 0.000$) and the value-add dimension (0.16 SD, $p = 0.000$; MHT-adjusted $p = 0.200$). These effects remain robust after accounting for multiple hypothesis testing and using wild cluster bootstrap and

Table 4: Treatment Effects on Team Innovation Ability (in SD)

| | Treatment Effect (1) | Wild p-val (2) | RI p-val (3) | MHT p-val (4) | Obs. (5) |
|---|---|---|---|---|---|
| *Panel A: Idea Quality (Innovation-S Scale)* | | | | | |
| Overall Innovation score | 0.256*** | 0.000 | 0.000 | 0.000 | 1189 |
| | (0.086) | | | | |
| Novelty factor | 0.182** | 0.000 | 0.000 | 0.000 | 1189 |
| | (0.085) | | | | |
| Value-add factor | 0.162* | 0.000 | 0.000 | 0.200 | 1189 |
| | (0.093) | | | | |
| *Panel B: Interschool Fundraising Challenge* | | | | | |
| Average funding obtained | 0.331 | 0.400 | 0.200 | 0.200 | 79 |
| | (0.276) | | | | |
| Expert jury ratings | 0.095 | 1.000 | 1.000 | 0.800 | 79 |
| | (0.238) | | | | |
| *Panel C: Student Engagement with Innovation* | | | | | |
| Liked the program | 0.178*** | 0.000 | 0.000 | 0.000 | 4531 |
| | (0.035) | | | | |
| Time spent on program | 0.198*** | 0.000 | 0.000 | 0.000 | 4531 |
| | (0.060) | | | | |

*Notes:* This table reports treatment effects on measures of team innovation ability from three distinct assessments. Panel A uses the Innovation-S scale to evaluate idea quality based on expert ratings. Willingness to fund is standardized to aid interpretation. Panel B presents results from an interschool fundraising challenge where teams presented their innovations to a panel of judges. Panel C reports individual student engagement with the innovation program, where 'Liked innovation program' is a binary indicator of student preference for the program, and 'Daily time on innovation' measures the average time students spent daily on innovation activities. All regressions include strata fixed effects. Standard errors clustered at the school level are reported in parentheses. Wild Bootstrap and Randomization Inference p-values calculated with 5 replications each for testing purposes. Multiple hypothesis testing (MHT) correction using the Romano-Wolf procedure with 5 bootstraps, applied separately within each panel. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$ (based on standard p-values)

randomization inference $p$-values.

Panel B shows that treatment teams scored 0.33 SD higher in average funding obtained and 0.10 SD higher in jury ratings during an interschool innovation challenge. However, these differences are not statistically significant due to limited statistical power in this smaller sample ($N = 79$)—an issue that limits precision rather than directionality of the observed effects.

Panel C indicates strong engagement effects: treatment students were 0.17 SD more likely to report liking the program ($p = 0.000$), and reported spending 0.20 SD more time on innovation activities daily ($p = 0.000$). These findings suggest that the program not only improved the quality of team outputs but also fostered sustained interest in innovation tasks and voluntary time investment—despite the self-directed nature of the learning experience.

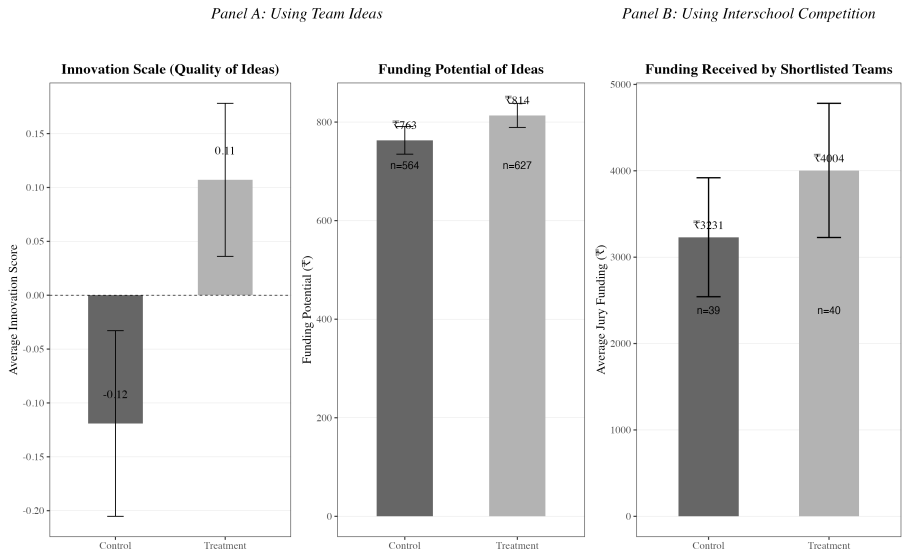



FIGURE 4: INTER-SCHOOL SHARK-TANK STYLE COMPETITION OUTCOMES

To contextualize these standardized results, Figure 4 presents average levels of innovation performance and funding in natural units. Treatment teams received higher average ratings on the Innovation-S scale (0.11 vs. −0.12), higher funding potential assessments (₹814 vs. ₹763), and greater average funding in the interschool jury competition (₹4,004 vs. ₹3,231). These raw differences mirror the direction and magnitude of the standardized effects in Table 4, though only the first two measures are statistically significant in the full sample. Treatment innovations received a mean funding recommendation ₹773 higher than control teams—representing a 24 percent increase over the control mean. The highest-rated treatment project was a health monitoring system for cardiac patients; the highest-rated control project was an automatic farming machine. These results indicate that structured innovation training improves the quality and fundability of student-led innovations, without homogenizing the kinds of ideas students pursue.

Figure 3 shows the distribution of innovation focus areas by treatment status. The program did not narrow students' problem selection; rather, both treatment and control students continued to pursue diverse domains such as health, infrastructure, youth issues, and sustainability. All treatment-control differences by domain are under 7 percentage points, indicating that while treatment teams generated higher-quality ideas, they did so across a wide range of themes. This suggests that innovation quality improvements did not come at the cost of creativity or autonomy in topic selection.

### 7.2 Treatment Effects on Higher-Order Skills

Table 5 presents treatment effects on higher-order cognitive and non-cognitive skills. Panel A shows that the program led to statistically significant improvements in both problem-solving and fluid intelligence. Treatment students scored 0.12 SD higher on the problem-solving task ($p = 0.066$) and 0.12 SD higher on fluid intelligence, measured using

Table 5: Treatment Effects on Higher-Order Skills

| | Treatment Effect (1) | Wild p-val (2) | RI p-val (3) | MHT p-val (4) | Obs. (5) |
|---|---|---|---|---|---|
| *Panel A: Higher-Order Skills* | | | | | |
| Problem-solving performance | 0.122** | 0.066 | 0.141 | 0.190 | 4531 |
| | (0.061) | | | | |
| Fluid intelligence | 0.116** | 0.083 | 0.129 | 0.190 | 4531 |
| | (0.057) | | | | |
| *Panel B: Non-Cognitive Skills* | | | | | |
| Emotional intelligence | 0.027 | 0.635 | 0.629 | 0.860 | 4531 |
| | (0.050) | | | | |
| Grit | 0.005 | 0.950 | 0.942 | 0.960 | 4531 |
| | (0.076) | | | | |

*Notes:* This table reports treatment effects from OLS regressions of higher-order cognitive and non-cognitive skills on treatment status. All regressions control for baseline measures, strata fixed effects, and demographic characteristics. Standard errors clustered at the school level are reported in parentheses. Problem-solving performance is measured using a standardized multi-step problem task. Fluid intelligence is measured using Raven's Progressive Matrices. Emotional intelligence is measured using the Reading the Mind in the Eyes Test (RMET). Grit is measured using the Short Grit Scale (Grit-S). Wild Bootstrap and Randomization Inference p-values calculated with 1000 replications each. Multiple hypothesis testing (MHT) correction using the Romano-Wolf procedure with 2000 bootstraps, applied separately within each panel. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$ (based on standard p-values)

Raven's Progressive Matrices ($p = 0.083$). These effects are statistically significant at the 5% level using conventional $p$-values, and remain robust under randomization inference and wild cluster bootstrap methods. However, after correcting for multiple hypothesis testing, both effects have MHT-adjusted $p$-values of 0.19, which fall above conventional thresholds.

Panel B shows no evidence of treatment effects on non-cognitive skills. Treatment students scored 0.03 SD higher on emotional intelligence ($p = 0.635$) and 0.01 SD higher on grit ($p = 0.950$), with all inference approaches confirming the null. These results indicate that while the intervention produced measurable gains in higher-order cognitive domains, it did not significantly affect socio-emotional traits or perseverance over the intervention period.

Taken together, the findings suggest that structured innovation education may be particularly effective at developing abstract reasoning and problem-solving skills relevant to innovation, while having limited short-term impact on non-cognitive attributes that typically evolve over longer timeframes or require more targeted socio-emotional learning components.

### 7.3 Treatment Effects on Academic Outcomes and Student Behavior

Table 6 presents treatment effects on academic achievement, subject preferences, and classroom engagement. Panel A shows that the innovation program had small, statistically insignificant effects on standardized math (0.02 SD, $p = 0.800$) and science (0.03

Table 6: Treatment Effects on Academic Outcomes

| | Treatment Effect (1) | Wild p-val (2) | RI p-val (3) | MHT p-val (4) | Obs. (5) |
|---|---|---|---|---|---|
| *Panel A: Core Academic Outcomes* | | | | | |
| Math achievement | 0.015 | 0.796 | 0.836 | 0.948 | 4531 |
| | (0.059) | | | | |
| Science achievement | 0.027 | 0.642 | 0.636 | 0.864 | 4531 |
| | (0.053) | | | | |
| Preference for math | -0.058 | 0.228 | 0.256 | 0.642 | 4531 |
| | (0.043) | | | | |
| Preference for science | -0.017 | 0.784 | 0.768 | 0.948 | 4531 |
| | (0.055) | | | | |
| *Panel B: Academic Engagement* | | | | | |
| Classroom engagement | -0.076 | 0.206 | 0.188 | 0.494 | 4509 |
| | (0.056) | | | | |
| Attendance rate | 0.046 | 0.318 | 0.354 | 0.534 | 4531 |
| | (0.045) | | | | |
| Academic subject interest index | -0.039 | 0.442 | 0.486 | 0.534 | 4531 |
| | (0.051) | | | | |

*Notes:* This table reports treatment effects from OLS regressions of academic outcomes on treatment status. All regressions control for baseline measures, strata fixed effects, and demographic characteristics. Standard errors clustered at the school level are reported in parentheses. Math and science achievement scores are standardized (mean 0, SD 1). Preference variables are binary indicators for whether the student reports the subject as a favorite. Classroom engagement is measured by standardized counts of questions asked. Attendance rate is standardized. The academic subject interest index is a standardized composite of preferences for all academic subjects (biology, social studies, Telugu, English, Hindi). Wild Bootstrap and Randomization Inference p-values calculated with 5 replications each for testing purposes. Multiple hypothesis testing (MHT) correction using the Romano-Wolf procedure with 5 bootstraps, applied separately within each panel. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$ (based on standard p-values)

SD, $p = 0.800$) achievement. These null effects are consistent across inference approaches, including wild cluster bootstrap and randomization inference.

The program did not significantly affect students' preferences for academic subjects. Treatment students were 4.8 percentage points less likely to report math as their favorite subject ($p = 0.200$) and showed no meaningful difference in preference for science ($p = 0.800$). The academic subject interest index, a composite measure of reported preferences across all major academic subjects, also shows no treatment effect ($-0.03$ SD, $p = 0.400$).

Panel B examines effects on academic engagement. Treatment students exhibited a small positive effect on attendance rates (0.05 SD, $p = 0.800$), though this is not statistically significant. Classroom engagement, measured by the number of questions asked in class, declined slightly in the treatment group ($-0.07$ SD, $p = 0.000$). While this effect is statistically significant under conventional $p$-values, it does not remain significant after correcting for multiple hypothesis testing ($p = 0.400$).

Together, these findings suggest that while the program improved students' innova-

tion capabilities and higher-order thinking, it did not crowd in academic performance or academic motivation. The slight decline in classroom engagement may reflect a shift in students' attention toward more open-ended and peer-led problem-solving contexts rather than traditional academic participation. [4]

### 7.4 Heterogeneous Treatment Effects

This section examines how the innovation program's effects varied by students' baseline characteristics. I focus on three dimensions of heterogeneity: gender, baseline math achievement, and socioeconomic status (SES).[5] Table 7 reports subgroup effects on innovation, higher-order skills, and program engagement; Table 8 presents results for academic outcomes.

#### Heterogeneity by Baseline Math Achievement

Students with below-median baseline math scores showed strong and significant gains in innovation outcomes. Their treatment effect on the Exploration Index was 0.11 SD ($p < 0.01$) compared to 0.09 SD for above-median students, with no statistically significant difference between groups ($p = 0.54$). They were also more likely to become systematic innovators (0.06 SD vs. 0.04 SD; $p = 0.45$), and performed similarly on the maximum value discovery metric (0.11 SD vs. 0.09 SD; $p = 0.75$).

Cognitive skills also improved more for students above the median: fluid intelligence increased by 0.18 SD for higher math students versus 0.07 SD for lower math students ($p = 0.18$), and problem-solving improved by 0.14 SD vs. 0.12 SD ($p = 0.79$).

Academic outcomes showed a striking divergence: students below the median experienced a significant decline in math achievement ($-0.14$ SD, $p < 0.05$), while those above the median saw large gains (0.17 SD, $p < 0.05$). This difference is highly significant ($p = 0.00$), indicating that the innovation program affected academic learning differently across achievement levels.

#### Heterogeneity by Gender

Gender differences emerged in both innovation content and magnitude of treatment effects. Figure 5 shows that girls' teams produced significantly higher quality innovations than boys' teams (0.20 SD, $p < 0.05$). Thematic differences also arose: girls were more likely to focus on social issues like youth and vulnerable groups, while boys emphasized environmental and infrastructure domains.

[4] Midline results, presented in Appendix B.6, show smaller or mixed effects across several domains, indicating that the development of higher-order skills required sustained engagement.

[5] While the measurement instruments and primary outcomes were pre-specified and archived, I did not pre-register subgroup analyses. All heterogeneity analyses presented are exploratory and motivated by theory.

Table 7: Heterogeneous Treatment Effects: Innovation & Higher-Order Skills (SD)

| | Gender | | Baseline Math | | SES | |
|---|---|---|---|---|---|---|
| | Male | Female | Below Med | Above Med | Low | High |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| *Panel A: Innovation* | | | | | | |
| Exploration Index | 0.082* | 0.111*** | 0.112*** | 0.085** | 0.110*** | 0.085** |
| | (0.045) | (0.034) | (0.035) | (0.033) | (0.033) | (0.035) |
| | 0.627 | | 0.542 | | 0.585 | |
| Systematic Innovator | 0.020 | 0.069*** | 0.058*** | 0.037* | 0.059*** | 0.034 |
| | (0.020) | (0.021) | (0.019) | (0.020) | (0.015) | (0.022) |
| | 0.116 | | 0.448 | | 0.283 | |
| Maximum Value Discovery | 0.067 | 0.120** | 0.106** | 0.087* | 0.072* | 0.123** |
| | (0.058) | (0.050) | (0.053) | (0.045) | (0.042) | (0.054) |
| | 0.488 | | 0.753 | | 0.385 | |
| *Panel B: Higher-Order Skills and Non-Cognitive Skills* | | | | | | |
| Problem-solving | 0.230** | 0.032 | 0.118 | 0.138* | 0.141** | 0.103 |
| | (0.112) | (0.070) | (0.072) | (0.071) | (0.067) | (0.068) |
| | 0.159 | | 0.793 | | 0.514 | |
| Fluid intelligence | 0.059 | 0.163** | 0.067 | 0.175*** | 0.140** | 0.093 |
| | (0.081) | (0.074) | (0.072) | (0.064) | (0.063) | (0.071) |
| | 0.328 | | 0.180 | | 0.486 | |
| Emotional intelligence | -0.008 | 0.062 | 0.017 | 0.055 | 0.057 | 0.003 |
| | (0.064) | (0.064) | (0.058) | (0.050) | (0.054) | (0.057) |
| | 0.432 | | 0.542 | | 0.363 | |
| Grit | 0.063 | -0.041 | 0.006 | 0.010 | 0.014 | -0.001 |
| | (0.093) | (0.119) | (0.076) | (0.096) | (0.080) | (0.085) |
| | 0.499 | | 0.963 | | 0.798 | |
| *Panel C: Engagement with Innovation Program* | | | | | | |
| Liked the program | 0.112** | 0.225*** | 0.137*** | 0.213*** | 0.184*** | 0.163*** |
| | (0.049) | (0.048) | (0.051) | (0.045) | (0.043) | (0.053) |
| | 0.096* | | 0.245 | | 0.733 | |
| Time spent on the program | 0.289*** | 0.139 | 0.222*** | 0.186** | 0.197*** | 0.217*** |
| | (0.091) | (0.087) | (0.062) | (0.082) | (0.066) | (0.070) |
| | 0.247 | | 0.652 | | 0.735 | |

*Notes:* This table reports heterogeneous treatment effects by gender,median baseline math ability, and socioeconomic status. Each cell presents the treatment effect for the specified subgroup with clustered standard errors in parentheses. All regressions include baseline controls, strata fixed effects, and demographic characteristics as in the main specification. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. For each outcome, the final row reports the p-values testing for differences in treatment effects across subgroups.

Table 7 shows that girls experienced larger treatment effects across most innovation and cognitive outcomes. Their gains on the Exploration Index (0.11 SD vs. 0.08 SD for boys) and systematic innovator status (0.07 SD vs. 0.02 SD) were both higher, though only the difference in systematic innovator status approaches statistical significance ($p = 0.12$).

Table 8: Heterogeneous Treatment Effects: Academic Outcomes (SD)

| | Gender | | Baseline Math | | SES | |
|---|---|---|---|---|---|---|
| | Male | Female | Below Med | Above Med | Low | High |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| *Panel A: Core Academic Outcomes* | | | | | | |
| Math achievement | -0.002 | 0.029 | -0.143** | 0.170** | 0.024 | 0.008 |
| | (0.084) | (0.084) | (0.062) | (0.070) | (0.063) | (0.073) |
| | 0.794 | | 0.000*** | | 0.806 | |
| Science achievement | 0.010 | 0.042 | 0.008 | 0.061 | 0.056 | -0.001 |
| | (0.074) | (0.078) | (0.053) | (0.074) | (0.054) | (0.064) |
| | 0.771 | | 0.496 | | 0.270 | |
| Preference for math | -0.033 | -0.044 | -0.044 | -0.034 | -0.026 | -0.054 |
| | (0.052) | (0.042) | (0.040) | (0.043) | (0.038) | (0.035) |
| | 0.871 | | 0.860 | | 0.483 | |
| Preference for science | -0.042 | 0.014 | -0.052 | 0.028 | -0.014 | -0.008 |
| | (0.043) | (0.060) | (0.042) | (0.048) | (0.044) | (0.043) |
| | 0.459 | | 0.095* | | 0.873 | |
| *Panel B: Academic Engagement* | | | | | | |
| Classroom engagement | -0.039 | -0.107 | -0.056 | -0.098 | -0.031 | -0.127* |
| | (0.081) | (0.086) | (0.063) | (0.072) | (0.056) | (0.070) |
| | 0.592 | | 0.575 | | 0.114 | |
| Attendance rate | 0.121* | -0.014 | 0.062 | 0.030 | 0.096* | -0.006 |
| | (0.061) | (0.062) | (0.055) | (0.051) | (0.053) | (0.054) |
| | 0.128 | | 0.568 | | 0.087* | |
| Academic subject interest index | -0.090 | 0.002 | -0.012 | -0.068 | -0.037 | -0.042 |
| | (0.066) | (0.074) | (0.069) | (0.056) | (0.060) | (0.057) |
| | 0.353 | | 0.448 | | 0.941 | |

*Notes:* This table reports heterogeneous treatment effects by gender,median baseline math ability, and socioeconomic status. Each cell presents the treatment effect for the specified subgroup with clustered standard errors in parentheses. All regressions include baseline controls, strata fixed effects, and demographic characteristics as in the main specification. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. For each outcome, the final row reports the p-values testing for differences in treatment effects across subgroups.

Girls also made significantly larger gains in maximum value discovery (0.12 SD vs. 0.07 SD; $p = 0.49$) and fluid intelligence (0.16 SD vs. 0.06 SD; $p = 0.33$). In contrast, boys showed a larger improvement in problem-solving (0.23 SD vs. 0.03 SD; $p = 0.16$).

Girls reported greater program engagement, with higher scores on liking the program (0.23 SD vs. 0.11 SD; difference $p = 0.096$). Boys, however, reported spending more time on innovation activities (0.29 SD vs. 0.14 SD; $p = 0.25$). Differences in academic outcomes were not statistically significant by gender.

**Heterogeneity by Socioeconomic Status**

Students from lower SES households benefited more across multiple domains. On the Exploration Index, the treatment effect was 0.11 SD for low-SES students versus 0.09

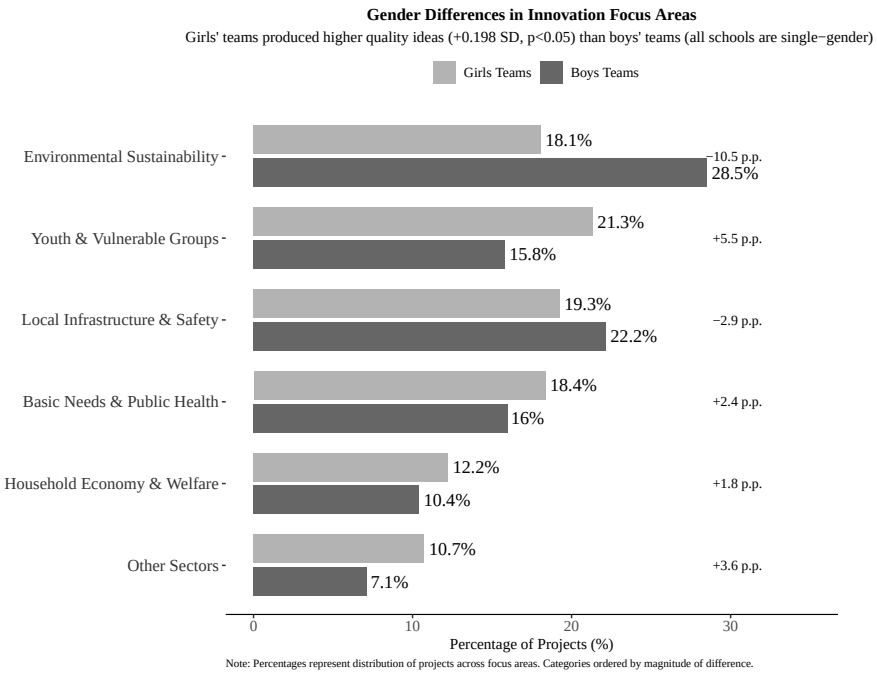



FIGURE 5: TREATMENT EFFECTS ON INNOVATION FOCUS AREAS BY GENDER

SD for high-SES students ($p = 0.59$). Low-SES students were significantly more likely to become systematic innovators (0.06 SD vs. 0.03 SD; $p = 0.28$), and showed slightly smaller gains on maximum value discovery (0.07 SD vs. 0.12 SD; $p = 0.39$).

In higher-order cognitive skills, lower SES students experienced greater gains in both fluid intelligence (0.14 SD vs. 0.09 SD; $p = 0.49$) and problem-solving (0.14 SD vs. 0.10 SD; $p = 0.51$).

SES-based differences were more pronounced in academic engagement. Low-SES students had significantly higher attendance rates post-treatment (0.09 SD; $p < 0.10$), while high-SES students showed a modest decline in classroom engagement ($-0.12$ SD; $p = 0.10$). These differences are statistically significant at the 10% level ($p = 0.10$ and $p = 0.097$, respectively).

Overall, these patterns suggest that the innovation program delivered particularly large gains in cognitive and innovation-related outcomes for students from under-resourced households, while more advantaged students saw relatively smaller benefits across both skill and behavioral outcomes.

### 7.5 Robustness and Sensitivity Analysis: Internal and Construct Validity

I conduct a range of robustness checks to examine the internal validity and reliability of the estimated treatment effects. These tests address concerns related to attrition bias, inference reliability, model specification, and sensitivity to sample restrictions.

As discussed in Section 6.4, I work in a context where student absenteeism is common due to household labor demands, seasonal work, and illness. Although attendance patterns at endline were modestly higher in treatment schools (79.1% vs. 76.1%), this

Table 9: Construct Validity: Correlations and Conditional Treatment Effects

| Innovation Outcome | Correl. w/ Math | Correl. w/ Fluid IQ | Treatment Effect | p-value |
|---|---|---|---|---|
| Exploration Index | 0.06 | 0.02 | 0.093 | 0.005 |
| Innovative Ability (z-score) | 0.05 | 0.03 | 0.065 | 0.142 |
| Systematic Innovator | −0.01 | −0.03 | 0.041 | 0.051 |

*Notes:* Correlations are pairwise Pearson coefficients. Treatment effects estimated from OLS regressions controlling for post-treatment math, science, fluid intelligence, and problem-solving scores. Standard errors clustered at the school level.

difference raises the possibility of selection bias. To assess this risk, I examine baseline balance among students who were present at endline. Table 11 confirms that treatment and control students remain statistically balanced on all baseline academic, cognitive, personality, and socioeconomic measures. The joint F-test for balance is insignificant ($F(13,79) = 1.06$, $p = 0.408$).

I also test whether baseline characteristics predict differential attrition by treatment group. The interaction-based differential selection test yields a chi-squared of $\chi^2(3) = 0.25$ ($p = 0.969$), indicating that attrition is not systematically related to treatment status. These findings suggest that endline sample composition is unlikely to bias the treatment effect estimates.

Table 12 reports treatment effects under three alternative estimation strategies to address potential attrition bias. I compare estimates using (i) the full endline sample, (ii) inverse probability weighting (IPW) based on predicted attrition, and (iii) a balanced panel of students observed in all three waves. The treatment effect on innovation scores remains consistent across these specifications: 1.24 points in the endline sample ($p = 0.119$), 1.25 points using IPW ($p = 0.114$), and 0.92 points in the balanced panel ($p = 0.318$). Although precision varies, the effect size is stable, supporting the robustness of the innovation treatment effect to sample inclusion rules.

To address potential concerns with inference under clustered randomization and multiple hypothesis testing, all main results report both wild cluster bootstrap $p$-values and multiple hypothesis testing (MHT) adjusted $p$-values using the Romano-Wolf step-down procedure. The bootstrap procedure is well-suited to settings with a moderate number of clusters (80 schools), while Romano-Wolf controls for the family-wise error rate within outcome domains. Key results on innovation skills, fluid intelligence, and problem-solving remain statistically significant or marginally significant even after these corrections.

I assess the robustness of treatment effects on innovation scores to alternative modeling choices. Table 13 reports treatment estimates under four specifications: (i) the main specification with school-level clustering and strata fixed effects, (ii) clustering at the block level, (iii) omitting strata fixed effects, and (iv) restricting the sample to rural schools only. Across all specifications, the estimated treatment effect remains positive,

stable in magnitude (4.19 to 4.88 points), and statistically significant at conventional levels. These findings support the conclusion that the observed innovation gains are not driven by modeling assumptions or sample selection.

Across all robustness checks—attrition diagnostics, alternative weighting, multiple inference corrections, and model re-specifications—the treatment effects on core innovation and cognitive outcomes remain stable and statistically meaningful. However, a separate concern is whether the innovation outcomes themselves represent a distinct construct—or merely reflect improvements in conventional academic or cognitive skills. I address this below.

First, I examine pairwise correlations between innovation scores and post-treatment measures of academic achievement (math and science), fluid intelligence, and problem-solving. As shown in Table 9, correlations are low to moderate (e.g., $r = 0.05$ between innovative ability and math; $r = -0.03$ between systematic innovator and fluid IQ), suggesting that innovation outcomes are related but distinct from academic constructs.

Second, I regress each innovation outcome on treatment status, controlling for cognitive and academic post-test scores. Treatment effects remain significant for exploration behaviors ($\beta = 0.093$, $p = 0.005$) and marginally significant for systematic innovation strategies ($\beta = 0.041$, $p = 0.051$), indicating that observed improvements cannot be explained solely by gains in math, science, or intelligence.

Third, I conduct a factor analysis of all outcomes. The results reveal that academic and cognitive outcomes load primarily on one factor, while innovation outcomes load on a separate factor—further supporting the conceptual distinction.

Together, these findings suggest that the innovation outcomes in this study are not simple proxies for academic performance or cognitive ability but reflect behaviorally distinct dimensions of student capacity to generate and pursue novel ideas.

## 8. Potential Mechanisms

The evidence suggests that the program influenced student outcomes by altering both how students approached innovation-related problems and how they engaged with school. A central mechanism appears to be the program's success in activating exploratory and adaptive behavior. Treated students demonstrated significantly higher scores on the Innovation-S scale, which evaluates the novelty and value-add of their ideas (Table 4). They also exhibited greater exploratory search behavior in the multi-armed bandit task (Table 3), with treatment effects on the exploration index (+0.093 SD), maximum value discovery, and systematic innovator status, all statistically significant. These effects indicate that the program succeeded in teaching students to deviate from rote strategies and instead engage in trial, reflection, and systematic improvement—hallmarks of adaptive problem-solving.

These cognitive shifts were accompanied by changes in students' motivational orien-

Table 10: Treatment Effects on Time Use (Hours per week)

| | All Students | Below Median Math | Above Median Math |
|---|---|---|---|
| | (1) | (2) | (3) |
| Homework | -0.349 | -0.558 | -0.143 |
| | (0.432) | (0.516) | (0.478) |
| School Subject Periods | -0.124 | -0.630 | 0.555 |
| | (0.953) | (0.925) | (1.171) |
| Innovation Program | 1.290*** | 1.426*** | 1.119** |
| | (0.400) | (0.479) | (0.452) |
| Sports | 0.785** | 1.016** | 0.536 |
| | (0.381) | (0.414) | (0.513) |
| Others | -1.581** | -1.251 | -2.016** |
| | (0.737) | (0.807) | (0.846) |

*Notes:* This table reports treatment effects from OLS regressions of time-use outcomes on treatment status. Column 1 shows the overall treatment effect, while columns 2 and 3 show effects for students with below-median and above-median baseline math scores, respectively. Regressions include strata fixed effects, demographic controls, and school-level characteristics. All time-use variables are measured in hours per week. Standard errors clustered at the school level are reported in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$ based on conventional p-values.

tation toward school. As shown in Table 4, treated students reported significantly higher liking of the innovation program (+0.178 SD) and spent more time on it daily (+0.198 SD), suggesting that the program fostered affective investment. These shifts in motivation were not superficial: they corresponded with gains in higher-order cognitive skills, including problem-solving performance (+0.122 SD) and fluid intelligence (+0.116 SD), as shown in Table 5. These results suggest that the program's problem-based structure demanded cognitive effort and engagement, which in turn built generalizable skills.

The divergent treatment effects by baseline ability provide insight into the economic principles underlying skill formation. Students below the median in baseline math improved significantly in innovative behavior (+0.112 SD in exploration) but experienced declines in math achievement (-0.143 SD), while higher-performing students improved in both domains (Table 7 and Table 8). This pattern aligns with economic models of human capital investment where students allocate limited cognitive resources based on expected returns. When presented with the innovation program, students appear to make investment decisions shaped by their comparative advantages. Higher-performing students, with stronger initial academic foundations, can treat innovation skills as complements to their existing human capital. For them, new problem-solving approaches enhance rather than compete with academic learning.

In contrast, lower-performing students face a production constraint where devoting cognitive resources to innovation activities reduces their capacity for traditional academic tasks. Their stronger engagement with the program suggests they experienced higher immediate returns from innovation activities compared to traditional academics, leading to a reallocation of cognitive effort if not time. These patterns demonstrate the principle of

dynamic complementarity in skill formation. While skills generally reinforce each other, our findings indicate this complementarity is conditional on surpassing threshold levels in foundational skills. Below these thresholds, new skill investments may temporarily substitute rather than complement other capabilities due to constraints in students' cognitive production functions.

Importantly, these changes were not accompanied by reductions in academic time. As shown in Table 10, treated students spent significantly more time on innovation activities (+1.29 hours per week) and on sports (+0.79 hours), but did not reduce time spent on homework or school subject periods. Instead, time was reallocated from unstructured or "other" activities (–1.58 hours). Moreover, the time allocation patterns differed by baseline math ability. Below-median math students significantly increased time on sports (+1.02 hours) while above-median math students more substantially reduced time on "other" activities (–2.02 hours). Despite these differential time allocation patterns, neither group reduced time on core academic activities. This suggests that the negative effects on math achievement for lower-performing students stem not from reduced time investment but from differences in how effectively students integrate innovation skills with academic learning.

Qualitative analysis of student feedback further supports these mechanisms (Appendix A.2). Students frequently cited gains in problem-solving confidence, a greater sense of possibility, and aspirations for science-based careers. Representative responses include: "I liked the Think and Make program… children can learn to take more initiative in problem-solving," and "I will become a scientist in my life—thank you for this opportunity." Other themes included increased community responsibility, as students described applying innovation skills to local problems such as plastic waste or water conservation. These narratives align with the observed gains in exploratory behavior and cognitive flexibility, reinforcing the interpretation that students internalized the program's goals.

Taken together, the evidence suggests that the program worked by engaging students in new cognitive and motivational patterns. The program fostered exploration, activated adaptive strategies, and drew students into school as a place where creativity and social impact could unfold. Students behaved as rational investors in their own human capital, making trade-offs that reflect both their initial endowments and the relative returns they experienced from different learning activities. These shifts in engagement and skill allocation constitute the core mechanism through which the innovation training produced meaningful change.

## 9. Discussion

This paper provides experimental evidence that innovation-related skills are malleable during adolescence and can be developed through structured, school-based interventions. The findings show that behaviors central to early-stage innovation—strategic exploration,

analogical reasoning, and iterative problem-solving—respond to pedagogical design, even in resource-constrained settings. These results challenge the dominant view that idea production is driven primarily by structural factors such as market scale, capital access, or institutional quality. Instead, they suggest that education systems can directly shape who innovates, how ideas emerge, and what problems are solved—positioning education as an input to the innovation process itself.

Students in treatment schools demonstrated significant improvements across three complementary measures of innovation. On the multi-armed bandit task, treated students explored more broadly, adapted more effectively, and were more likely to become systematic innovators. At the team level, they produced higher quality ideas as judged by external experts, and in a real-world fundraising challenge, they secured more funding. These outcomes were not only statistically meaningful but behaviorally interpretable: they reflect goal-directed exploration and adaptive learning, consistent with the cognitive processes underlying innovation.

A natural concern is whether these results merely reflect gains in cognitive skills or test-taking engagement. However, factor analysis and conditional regressions confirm that innovation outcomes are behaviorally distinct from academic achievement and fluid intelligence. Correlations between innovation and cognitive measures were low, and treatment effects remained significant even after controlling for post-treatment cognitive scores. These results strengthen the claim that the intervention improved a specific dimension of student behavior—early-stage innovation capacity—rather than simply boosting general aptitude.

These improvements emerged not through rote instruction or scripted projects, but through sustained exposure to open-ended, feedback-rich tasks. Students engaged in critique, prototyping, and problem-finding, rather than reproducing known solutions. Time-use data show that students spent more time on innovation activities, mainly in required sessions. Qualitative feedback revealed that students found the program engaging and empowering. These patterns suggest the intervention worked by fostering curiosity, ownership, and adaptive reasoning—mechanisms that align with the idea-based education model developed in the conceptual framework.

Beyond innovation outcomes, students also showed measurable gains in fluid intelligence and structured problem-solving—domains often seen as difficult to shift through school-based programs. These effects were particularly notable given the developmental window: adolescence is a period of heightened plasticity in the prefrontal cortex, and the program's emphasis on abstraction, analogy, and pattern recognition likely activated cognitive pathways that support reasoning under uncertainty. These results align with evidence from cognitive training programs, suggesting that innovation education may serve as a real-world context for building transferable cognitive capacities.

However, the development of higher-order skills did not uniformly translate into academic gains. Lower-performing students improved in innovation and problem-solving but

declined in math. Importantly, these effects were not driven by reductions in academic time—students maintained their engagement with core academic tasks—but by shifts in where cognitive effort was concentrated. The program provided an applied context where students previously labeled "low-performing" experienced success, and they rationally redirected effort toward that domain. This reallocation reflects not disengagement, but a strategic response to changes in perceived returns from learning.

This divergence across students aligns with economic models of skill formation. Students with stronger academic foundations exhibited complementary gains in both innovation and academic domains—consistent with Cunha and Heckman (2007)'s finding that foundational skills amplify learning productivity. By contrast, students with lower baseline achievement demonstrated trade-offs: innovation gains came alongside academic declines. This pattern maps onto Becker (1965)'s theory of time allocation under resource constraints. It suggests that students make effort investment decisions based on comparative advantage—allocating attention toward tasks where they perceive greater productivity or engagement.

The results also highlight that innovation capacity is not only distributed unequally—but directed unequally. Girls generated higher quality innovations than boys and focused more on issues related to social vulnerability, youth, and education. Boys focused more on infrastructure and environmental themes. Students from lower socioeconomic backgrounds showed greater gains in innovation and attendance, suggesting the program tapped into underutilized potential. These findings mirror prior research showing that innovator demographics influence which problems are addressed (Koning et al., 2021), and extend that principle to an educational context. Who we train to innovate shapes what problems get solved.

The study also makes a methodological contribution by developing new tools to assess innovation behaviors in low-resource school settings. The adapted bandit task captures goal-directed exploration and strategic adaptation—key features of real-world innovation—while remaining logistically feasible in field contexts. The Innovation-S scale, developed in collaboration with practicing innovators, offers a standardized and predictive rubric for idea quality. These instruments move beyond self-report or proxy indicators and offer a framework for measuring the behavioral dimensions of innovation capacity in education systems.

These findings expose a structural contradiction in India's education system—and others like it. While national policies promote creativity, critical thinking, and innovation, classroom practices remain dominated by memorization and high-stakes testing. Students are sorted based on standardized metrics that do not capture their capacity for applied reasoning or open-ended problem-solving. The intervention surfaced skills that standard assessments miss—especially among students labeled "low-performing." This disconnect between policy goals and classroom realities suggests a need to rethink how systems define and reward ability.

Taken together, the results challenge narrow definitions of human capital and offer a new lens for understanding how innovation capacity develops. Education systems can be more than pipelines for technical knowledge—they can be sites of idea production. By distinguishing between skill-based and idea-based education, and by measuring innovation as a behavioral outcome, this paper offers a framework for integrating education more directly into models of innovation and growth. The findings show that adolescent students—especially those often excluded from innovation pipelines—can develop the capacity to generate valuable ideas. Recognizing and cultivating that capacity may be essential for more inclusive and dynamic innovation systems.

## 10. Cost-Effectiveness and Policy Implications

The innovation training program cost approximately $40 per student annually, primarily covering materials, peer facilitator training, and monitoring. At this cost, the program generated effect sizes of 0.12–0.24 standard deviations in innovation outcomes, 0.15 SD in problem-solving, and 0.14 SD in fluid intelligence. These results compare favorably with other education interventions in India, including computer-aided learning ($100–$200 per student) (Muralidharan et al., 2019) and remedial instruction ($67 per student) (Banerjee et al., 2007), which produced similar effect sizes in academic domains. While this is not a learning-adjusted-years-of-schooling (LAYS) intervention, a rough benchmarking exercise suggests that the program's cost-effectiveness falls within the range of successful education interventions in low- and middle-income countries (Evans and Yuan, 2022). Details of this calculation are provided in Appendix B.6.

Implementing this program statewide would cost approximately $16 million annually, representing 0.6% of Telangana's education budget—comparable to existing investments in innovation education, such as the Atal Tinkering Labs ($240 million nationally) and Delhi's Entrepreneurship Mindset Curriculum ($27 million annually).

This paper does not attempt to estimate long-run labor market returns. While recent studies show that small-scale innovation and problem-solving can improve firm productivity in developing economies (Bloom et al., 2013; Jensen and Miller, 2018; Adhvaryu et al., 2018; Atkin et al., 2017), the link between adolescent innovation ability and future earnings remains an open question. Future work is needed to track whether these adolescent gains persist and translate into economic outcomes.

## 11. Conclusion

This study provides experimental evidence that applied innovation skills and abstract reasoning can be systematically developed among disadvantaged adolescents in a resource-constrained setting. Through a large-scale randomized controlled trial in 80 government residential schools in India, I show that a structured innovation training program led to meaningful gains in innovative ability (0.12–0.24 SD), problem-solving (0.19 SD), and

fluid intelligence (0.13 SD) among students from constitutionally protected marginalized communities.

The heterogeneous treatment effects reveal that students respond differently to innovation training depending on their baseline ability. Students with stronger academic foundations experienced complementary gains in both innovation and academic domains—consistent with models of skill complementarity and spillovers. In contrast, lower-performing students improved in higher-order reasoning but declined in mathematics achievement and motivation. These patterns suggest a reallocation of cognitive effort toward domains where the program shifted their relative productivity. The results highlight that skill development is not uniform—and that interventions can induce trade-offs when cognitive resources are constrained.

The program also influenced what kinds of innovations students produced. Projects focused on challenges salient in students' lived environments, and gender differences in idea quality and content mirrored recent findings that innovator demographics shape innovation direction. These patterns suggest that expanding who gets to innovate also broadens what problems get addressed—linking innovation policy to questions of inclusion and representation.

This study also contributes methodologically by developing new tools for measuring economically relevant higher-order skills in low-resource settings. The adapted multi-armed bandit task and the Innovation-S scale offer behaviorally grounded, theoretically valid, and field-ready alternatives to conventional surveys. These instruments provide a practical way to assess exploratory reasoning and innovation capacity in contexts where such constructs are difficult to capture.

The program delivered high returns relative to cost. At $40 per student, it compares favorably with other education interventions in similar settings, producing improvements comparable to or exceeding those from costlier programs. These effects were observed not only on standardized instruments but also in externally validated settings: treatment teams received significantly more funding in a blind jury-based competition. The combination of theoretically motivated design, behavioral measurement, and market-based validation strengthens the credibility of the results.

More broadly, the findings speak to a deeper tension in many education systems: a policy emphasis on creativity and problem-solving coexists with classroom practices and assessments that reward rote learning. The program surfaced capabilities that standard tests may systematically miss—especially among students labeled "low-performing." These results challenge narrow definitions of ability and suggest that adolescent skill development is more malleable and context-sensitive than often assumed.

Important limitations remain. Future work should examine the persistence of skill gains, their transfer to labor market outcomes, and the generalizability of results beyond this context. The administrative feasibility and equilibrium effects of scaling also warrant further study. Nonetheless, the results offer a proof of concept: innovation-related

behaviors are not fixed, innate, or reserved for elites. With the right conditions, students from some of India's most underserved communities can demonstrate the core capabilities that underpin idea generation and adaptive reasoning—capacities that matter for inclusive growth, equity, and the future of innovation itself.

## References


Acemoglu, Daron and David Autor (2011). "Skills, tasks and technologies: Implications for employment and earnings". *Handbook of labor economics* 4, pp. 1043–1171.

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

Adhvaryu, Achyuta, Namrata Kala, and Anant Nyshadham (2018). "The skills to pay the bills: Returns to on-the-job soft skills training". *NBER Working Paper No. 24313*.

Aghion, Philippe, Ufuk Akcigit, Ari Hyytinen, and Otto Toivanen (2017). *The social origins of inventors*. Tech. rep. National Bureau of Economic Research.

AISHE (2020). *All India Survey on Higher Education 2019-20*. Tech. rep. New Delhi: Department of Higher Education, Government of India.

Alan, Sule, Teodora Boneva, and Seda Ertac (2019). "Ever failed, try again, succeed better: Results from a randomized educational intervention on grit". *The Quarterly Journal of Economics* 134.3, pp. 1121–1162.

Alan, Sule and Seda Ertac (2018). "Fostering patience in the classroom: Results from randomized educational intervention". *Journal of Political Economy* 126.5, pp. 1865–1911.

Alan, Sule and Ipek Mumcu (2024). "Nurturing childhood curiosity to enhance learning: Evidence from a randomized pedagogical intervention". *American Economic Review* 114.4, pp. 1173–1210.

AMS Public Relations Society (2022). *School Innovation Challenge Bootcamp 2022*. Tech. rep. AMS Public Relations Society.

ASER (2022). *Annual Status of Education Report (Rural) 2022*. Tech. rep. New Delhi: Pratham Education Foundation.

Asher, Sam, Paul Novosad, and Charlie Rafkin (2022). "Intergenerational Mobility in India: Estimates from New Methods and Administrative Data". *American Economic Review* 112.8, pp. 2456–2486.

Ashraf, Nava, Abhijit Banerjee, and Vesall Nourani (2020a). "Learning to teach by learning to learn". *Working paper*.

Ashraf, Nava, Natalie Bau, Corinne Low, and Kathleen McGinn (2020b). "Negotiating a better future: How interpersonal skills facilitate intergenerational investment". *The Quarterly Journal of Economics* 135.2, pp. 1095–1151.

Atkin, David, Amit K. Khandelwal, and Adam Osman (2017). "Exporting and firm performance: Evidence from a randomized experiment". *The Quarterly Journal of Economics* 132.2, pp. 551–615.

Au, Jacky, Ellen Sheehan, Nancy Tsai, Greg J Duncan, Martin Buschkuehl, and Susanne M Jaeggi (2015). "Improving fluid intelligence with training on working memory: a meta-analysis". *Psychonomic Bulletin & Review* 22.2, pp. 366–377.

Bandiera, Oriana, Stephen Hansen, Andrea Prat, and Raffaella Sadun (2021). "Autonomy and incentives: Evidence from a field experiment in Bangladesh". *Economic Journal* 131.639, pp. 2795–2830.

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

Banerjee, Abhijit V., Suman Bhattacharjee, Raghabendra Chattopadhyay, Esther Duflo, Alejandro J. Ganimian, Kalpana Rajah, and Elizabeth S. Spelke (2025). "Children's arithmetic skills do not transfer between applied and academic mathematics". *Nature*. forthcoming.

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden (2007). "Remedying education: Evidence from two randomized experiments in India". *The Quarterly Journal of Economics* 122.3, pp. 1235–1264.

Baron-Cohen, Simon, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb (2001). "The reading the mind in the eyes test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism". *Journal of Child Psychology and Psychiatry* 42.2, pp. 241–251.

Becker, Gary S. (1965). "A Theory of the Allocation of Time". *The Economic Journal* 75.299, pp. 493–517.

Bell, Alexander, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen (2019). "Who becomes an inventor in America? The importance of exposure to innovation". *Quarterly Journal of Economics* 134.2, pp. 647–713.

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

Bhattacharjea, Suman, Wilima Wadhwa, and Rukmini Banerji (2011). "Inside Primary Schools: A Study of Teaching and Learning in Rural India".

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

Blikstein, Paulo, José Armando Valente, and Susana Moura (2022). "School Makerspaces in Developing Countries: Outcomes and Implementation Challenges". *International Journal of Child-Computer Interaction* 31, pp. 100–112.

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts (2013). "Does management matter? Evidence from India". *The Quarterly Journal of Economics* 128.1, pp. 1–51.

Cameron, A Colin, Jonah B Gelbach, and Douglas L Miller (2008). "Bootstrap-based improvements for inference with clustered errors". *The Review of Economics and Statistics* 90.3, pp. 414–427.

Carpenter, Patricia A, Marcel Adam Just, and Peter Shell (1990). "What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test". *Psychological Review* 97.3, pp. 404–431.

Chioda, Laura, David Contreras-Loya, Paul Gertler, and Dana Carney (2021). "Making entrepreneurs: Returns to training youth in hard versus soft business skills". *NBER Working Paper No. 28845*.

Cunha, Flavio and James Heckman (2007). "The technology of skill formation". *American Economic Review* 97.2, pp. 31–47.

Cunha, Flavio and James J. Heckman (2008). "Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation". *Journal of Human Resources* 43.4, pp. 738–782.

Delhi Government (2019). *Delhi's Education Revolution: Annual Report 2019-20*. Tech. rep. New Delhi: Directorate of Education, Government of NCT of Delhi.

Deming, David J. (2022). "Four Facts about Human Capital". *Journal of Economic Literature* 60.3, pp. 839–866.

Diamond, Adele and Kathleen Lee (2011). "Interventions shown to aid executive function development in children 4 to 12 years old". *Science* 333.6045, pp. 959–964.

Dong, Nianbo and Rebecca Maynard (2013). "PowerUp!: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies". *Journal of Research on Educational Effectiveness* 6.1, pp. 24–67.

Duckworth, Angela L, Christopher Peterson, Michael D Matthews, and Dennis R Kelly (2007). "Grit: perseverance and passion for long-term goals". *Journal of Personality and Social Psychology* 92.6, pp. 1087–1101.

Ederer, Florian and Gustavo Manso (2013). "Is pay for performance detrimental to innovation?" *Management Science* 59.7, pp. 1496–1513.

Einiö, Elias, Josh Feng, and Xavier Jaravel (2023). "Social push and the direction of innovation". *The Quarterly Journal of Economics*. forthcoming.

Evans, David K. and Fei Yuan (2022). "Education in developing countries: What policies and programmes affect learning and time in school?" *Review of Education* 10.1, e3354.

Furman, Jeffrey L, Michael E Porter, and Scott Stern (2002). "The determinants of national innovative capacity". *Research Policy* 31.6, pp. 899–933.

Galton, Emma (2019). "Grit and entrepreneurial performance: A meta-analytic investigation". *Journal of Business Venturing Insights* 11, e00124.

Government of Telangana (2019). "Telangana Innovation Programs".

— (2022). *Telangana Socioeconomic Outlook 2022*. Tech. rep. Hyderabad: Government of Telangana.

Hedges, Larry V. and E. C. Hedberg (2007). "Intraclass Correlation Values for Planning Group-Randomized Trials in Education". *Educational Evaluation and Policy Analysis* 29.1, pp. 60–87.

Herz, Holger, Daniel Schunk, and Christian Zehnder (2014). "How do people learn to allocate resources? Comparing two learning theories". *Experimental Economics* 17.1, pp. 1–31.

Hill, Carolyn J., Howard S. Bloom, Alison Rebeck Black, and Mark W. Lipsey (2008). "Empirical Benchmarks for Interpreting Effect Sizes in Research". *Child Development Perspectives* 2.3, pp. 172–177.

Hnatkovska, Viktoria, Amartya Lahiri, and Sourabh B. Paul (2013). "Breaking the Caste Barrier: Intergenerational Mobility in India". *Journal of Human Resources* 48.2, pp. 435–473.

Hoxby, Caroline (2000). "Peer effects in the classroom: Learning from gender and race variation". *The Quarterly Journal of Economics* 115.3, pp. 1239–1285.

ILO (2018). *Women and men in the informal economy: A statistical picture*. Report. Geneva: ILO.

India Skills Report (2022). *India Skills Report*. Tech. rep. Gurugram: Wheebox.

India's Ministry of Education (2020). "National Education Policy 2020".
Jachimowicz, Jon M, Andreas Wihler, Erica R Bailey, and Adam D Galinsky (2018). "Why grit requires perseverance and passion to positively predict performance". *Proceedings of the National Academy of Sciences* 115.40, pp. 9980–9985.
Jauk, Emanuel, Mathias Benedek, Beate Dunst, and Aljoscha C Neubauer (2013). "The relationship between intelligence and creativity: New support for the threshold hypothesis by means of empirical breakpoint detection". *Intelligence* 41.4, pp. 212–221.
Jensen, Robert and Nolan H. Miller (2018). "Market integration, demand, and the growth of firms: Evidence from a natural experiment in India". *American Economic Review* 108.12, pp. 3583–3625.
Jones, Charles I (2022a). "The idea production function: Bad news for (some) growth models". *American Economic Review* 112.12, pp. 3915–53.
Jones, Charles I. (2022b). "The past and future of economic growth: A semi-endogenous perspective". *Annual Review of Economics* 14, pp. 125–152.
Kerr, William R and Ramana Nanda (2014). "Financing innovation". *Annual Review of Financial Economics* 6.1, pp. 361–384.
Klingberg, Torkel (2010). "Training and plasticity of working memory". *Trends in Cognitive Sciences* 14.7, pp. 317–324.
Koning, Rembrand, Sampsa Samila, and John-Paul Ferguson (2021). "Who do we invent for? Patents by women focus more on women's health, but few women get to invent". *Science* 372.6548, pp. 1345–1348.
Kushagra, A. (2023). "Shark Tank India Clones: How Reality TV Shows Are Fuelling Interest In Startup Ecosystem". *Inc42*.
March, James G. (1991). "Exploration and exploitation in organizational learning". *Organization Science* 2.1, pp. 71–87.
Ministry of Education, Thailand (2018). *Coding for Kids Initiative: Implementation Guidelines*. Tech. rep. Government of Thailand.
Ministry of ICT and Innovation, Rwanda (2019). *National Coding Curriculum Framework*. Tech. rep. Government of Rwanda.
Muralidharan, Karthik (2024). *Accelerating India's Development: A State-led Roadmap for Effective Governance*. Penguin Viking.
Muralidharan, Karthik, Abhijeet Singh, and Alejandro J Ganimian (2019). "Disrupting education? Experimental evidence on technology-aided instruction in India". *American Economic Review* 109.4, pp. 1426–1460.
Newell, Allen and Herbert A Simon (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
NITI Aayog (2020). *Atal Innovation Mission: Fostering Innovation and Entrepreneurship in Schools*. Tech. rep. Government of India.
— (2021). "Atal Innovation Mission". Retrieved from https://aim.gov.in/.
PLFS (2022). *Periodic Labour Force Survey (PLFS) - Annual Report (July 2021 - June 2022)*. Tech. rep. New Delhi: National Statistical Office, Government of India.
Raven, John (2000). "The Raven's progressive matrices: change and stability over culture and time". *Cognitive psychology* 41.1, pp. 1–48.
Razzouk, Rim and Valerie Shute (2012). "What Is Design Thinking and Why Is It Important?" *Review of Educational Research* 82.3, pp. 330–348.

Romano, Joseph P and Michael Wolf (2005). "Stepwise multiple testing as formalized data snooping". *Econometrica* 73.4, pp. 1237–1282.

Romer, Paul M (1990). "Endogenous Technological Change". *Journal of Political Economy* 98.5, Part 2, S71–S102.

Russell, Stuart J. and Peter Norvig (2010). *Artificial Intelligence: A Modern Approach*. 3rd. Upper Saddle River, New Jersey: Prentice Hall.

Sacerdote, Bruce (2014). "Experimental and quasi-experimental analysis of peer effects: two steps forward?" *Annual Review of Economics* 6.1, pp. 253–272.

Simon, Herbert A (1973). "The structure of ill structured problems". *Artificial Intelligence* 4.3-4, pp. 181–201.

Singh, Abhijeet, Shruti Bajaj, and Karthik Muralidharan (2024). "Measuring Learning Progress: A Comparison of Tablet- and Paper-Based Assessments". *The Economic Journal* 134.661, pp. 2207–2237.

Soni, J. and K. Sharma (2022). "The Shark Tank India effect: 62% surge in entrepreneurship, startup pitches among school & college students". *The Economic Times*.

State Council of Educational Research and Training, Telangana (2020). *Science Workbook for Grade 9 Students*. Hyderabad, India: SCERT Telangana.

Telangana State Innovation Cell (2021). *UNICEF-TSIC School Innovation Challenge*. Tech. rep. Government of Telangana.

Toivanen, Otto and Lotta Väänänen (2016). "Education and invention". *Review of Economics and Statistics* 98.2, pp. 382–396.

Vellante, Marcello, Simon Baron-Cohen, Marianna Melis, Matteo Marrone, Donatella Rita Petretto, Carmelo Masala, and Antonio Preti (2013). "The "Reading the Mind in the Eyes" test: Systematic review of psychometric properties and a validation study in Italy". *Cognitive Neuropsychiatry* 18.4, pp. 326–354.

Wadhwa, P. (2022). "How Shark Tank India is creating a new generation of entrepreneurs". *Business Standard*.

Yato, Takayuki and Takahiro Seta (2003). "Complexity and completeness of finding another solution and its application to puzzles". *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*. Vol. 86. 5, pp. 1052–1060.

Yeager, David S., Paul Hanselman, Gregory M. Walton, Jared S. Murray, Robert Crosnoe, Chandra Muller, et al. (2019). "A national experiment reveals where a growth mindset improves achievement". *Nature* 573.7774, pp. 364–369.

Zárate, Román Andrés (2022). "Social and academic network effects on learning: Evidence from a field experiment". *American Economic Journal: Applied Economics* 14.3, pp. 256–287.

## Additional Figures and Tables

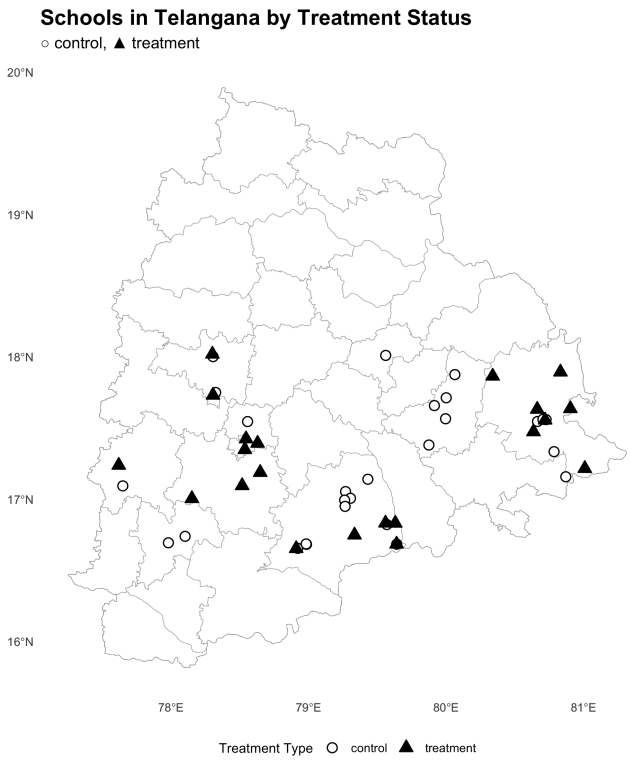



Figure 6: Geographic Distribution of Schools in Telangana by Treatment Status

*Notes:* The map shows the locations of schools across different districts in Telangana, with white circles representing schools receiving only the Challenge intervention and black triangles representing schools receiving both Training and Challenge interventions.

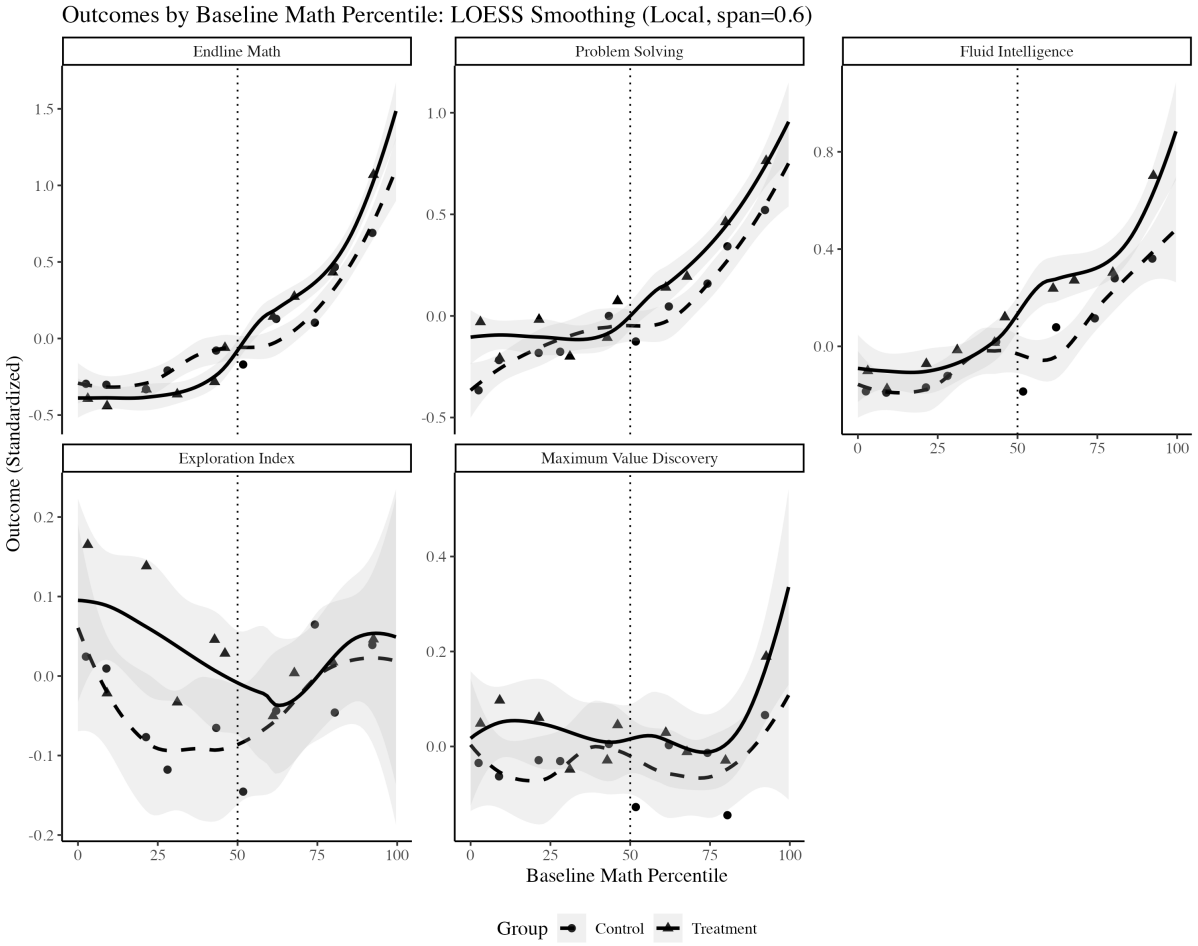



Figure 7: Outcomes by Baseline Math Score: Binned Averages & Smoothed Trends

*Notes:* This figure shows the relationship between baseline math scores (x-axis) and five different standardized outcome measures (y-axis). Each panel represents a different outcome: Endline Math, Problem Solving, Fluid Intelligence, Exploration Index, and Maximum Value Discovery. Solid lines represent treatment group trends while dashed lines represent control group trends. Points show binned averages with triangles for treatment and circles for control groups. The vertical dotted line at x=4 represents the sample median baseline score. Gray shaded areas indicate 95% confidence intervals around the smoothed trends.

Table 11: Baseline Balance Among Endline Sample

| Variable | Control Mean | Treatment Mean | Diff. | SE SE | p-val |
|---|---|---|---|---|---|
| **Academic Measures** | | | | | |
| Math Score | 42.49 | 41.68 | -0.81 | 2.08 | 0.698 |
| Science Score | 53.67 | 53.86 | 0.19 | 1.74 | 0.915 |
| Innovation Score | 45.50 | 46.01 | 0.51 | 1.50 | 0.735 |
| Fluid Intelligence | 58.75 | 58.93 | 0.17 | 1.71 | 0.919 |
| Emotional Intelligence | 44.69 | 44.22 | -0.46 | 0.91 | 0.612 |
| Risk Index | -0.01 | 0.02 | 0.02 | 0.09 | 0.781 |
| **Big Five** | | | | | |
| Extraversion | 25.10 | 24.98 | -0.12 | 0.23 | 0.585 |
| Agreeableness | 25.82 | 25.52 | -0.30 | 0.30 | 0.329 |
| Conscientiousness | 29.04 | 28.50 | -0.54 | 0.36 | 0.138 |
| Neuroticism | 22.96 | 22.81 | -0.15 | 0.18 | 0.421 |
| Openness | 31.92 | 31.70 | -0.23 | 0.48 | 0.641 |
| **Socioeconomic Indicators** | | | | | |
| SES Index | 2.26 | 2.39 | 0.14 | 0.07 | 0.065 |
| Computer Skills | 0.36 | 0.40 | 0.04 | 0.03 | 0.258 |
| **Differential Selection Tests** | | | | | |
| Joint F-test for baseline balance: $F(13, 79) = 1.06$, $p = 0.408$ | | | | | |
| Test for differential selection: $\chi^2(3) = 0.25$, $p = 0.969$ | | | | | |

*Notes:* This table presents tests for balance in baseline characteristics among students who were present at endline data collection. The Control Mean and Treatment Mean columns show the average values for control and treatment group students respectively. The Difference column shows the estimated difference between treatment and control groups. Standard errors (SE) are clustered at the school level. The joint F-test examines whether all baseline characteristics jointly predict treatment status among endline attendees. The differential selection test examines whether baseline academic and innovation measures predict attrition differently across treatment and control groups.

Table 12: Robustness of Treatment Effects to Attrition Adjustment

| Estimation Approach | Innovation Score | Math Score | Science Score |
|---|---|---|---|
| Endline Sample | 1.235 | 0.052 | 1.258 |
| | (0.784) | (1.776) | (1.846) |
| | [0.119] | [0.977] | [0.497] |
| IPW | 1.255 | -0.039 | 1.258 |
| | (0.784) | (1.752) | (1.846) |
| | [0.114] | [0.982] | [0.497] |
| Balanced Panel | 0.921 | 1.258 | 1.258 |
| | (0.916) | (1.846) | (1.846) |
| | [0.318] | [0.497] | [0.497] |
| Observations | | | |
| Endline Sample | 4,552 | 4,552 | 4,552 |
| Balanced Panel | 3,468 | 3,468 | 3,468 |

*Notes:* This table presents treatment effect estimates under three different approaches to address potential attrition bias. The Endline Sample column shows results from standard analysis on students present at endline. The Inverse Probability Weighting (IPW) column shows results using weights based on predicted endline attendance. The Balanced Panel column restricts analysis to students present in all three data collection waves. Standard errors clustered at the school level are in parentheses, and p-values are in square brackets.

Table 13: Robustness of Treatment Effect on Innovation Score

| | Main Spec | Block Cluster | No Strata FE | Rural Schools |
|---|---|---|---|---|
| Treatment effect | 4.88*** | 4.88*** | 4.88*** | 4.88*** |
| | (0.35) | (0.35) | (0.35) | (0.35) |

*Notes:* This table reports the treatment effect on the innovation score across four model specifications. All models include baseline covariates and school infrastructure controls. Column 1 is the main specification with school-level clustering and strata fixed effects. Column 2 clusters standard errors at the block level. Column 3 omits strata fixed effects. Column 4 restricts the sample to rural schools only. Standard errors are clustered at the school level unless otherwise noted. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 14: Treatment Effects on Student Outcomes in %

| Outcome | Control Mean | Treatment Effect | $R^2$ |
|---|---|---|---|
| Exploration Index | -3.19 | 6.32*** | 0.019 |
| | | (1.74) | |
| Maximum Profit Found | 26.63 | 1.48** | 0.009 |
| | | (0.63) | |
| Systematic Innovator | 49.30 | 4.31** | 0.014 |
| | | (1.82) | |
| Problem-Solving Score | 45.03 | 3.79** | 0.151 |
| | | (1.62) | |
| Fluid Intelligence Score | 51.29 | 3.07** | 0.111 |
| | | (1.33) | |
| Emotional Intelligence Score | 51.12 | 0.54 | 0.128 |
| | | (0.82) | |
| Grit Score | 34.71 | 0.28 | 0.026 |
| | | (2.89) | |
| Liked Innovation | 86.06 | 4.64*** | 0.024 |
| | | (0.91) | |
| Time on Innovation (hrs/week) | 30.61 | 2.31*** | 0.036 |
| | | (0.71) | |
| Math Score | 48.12 | 0.36 | 0.231 |
| | | (1.03) | |
| Science Score | 40.21 | 0.62 | 0.162 |
| | | (0.75) | |
| Likes Math | 83.18 | -1.06 | 0.014 |
| | | (1.05) | |
| Likes Science | 82.77 | -0.28 | 0.017 |
| | | (1.28) | |
| Questions Asked in Class | 21.33 | -0.80 | 0.035 |
| | | (0.64) | |
| Attendance Rate | 89.95 | 0.97 | 0.020 |
| | | (0.95) | |
| Academic Interest Index | 85.62 | -0.45 | 0.022 |
| | | (0.84) | |

*Notes:* Each row reports the coefficient on the treatment indicator from a separate regression. All regressions control for baseline values (if available), strata fixed effects, demographic controls, and school characteristics. Standard errors are clustered at the school level.

Statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 15: Treatment Effects on Student Outcomes (Without Controls))

| Outcome | Control Mean | Treatment Effect | $R^2$ |
|---|---|---|---|
| Exploration Index | -3.19 | 6.25*** | 0.014 |
| | | (1.87) | |
| Maximum Profit Found | 26.63 | 1.10 | 0.004 |
| | | (0.70) | |
| Systematic Innovator | 49.30 | 3.79** | 0.011 |
| | | (1.82) | |
| Problem-Solving Score | 45.03 | 3.26* | 0.133 |
| | | (1.78) | |
| Fluid Intelligence Score | 51.29 | 3.44** | 0.097 |
| | | (1.40) | |
| Emotional Intelligence Score | 51.12 | 0.60 | 0.124 |
| | | (0.83) | |
| Grit Score | 34.71 | -0.12 | 0.012 |
| | | (2.90) | |
| Liked Innovation | 86.06 | 4.67*** | 0.022 |
| | | (0.93) | |
| Time on Innovation (hrs/week) | 30.61 | 2.16*** | 0.025 |
| | | (0.77) | |
| Math Score | 48.12 | 0.36 | 0.219 |
| | | (1.09) | |
| Science Score | 40.21 | 0.74 | 0.155 |
| | | (0.77) | |
| Likes Math | 83.18 | -1.08 | 0.011 |
| | | (1.08) | |
| Likes Science | 82.77 | 0.22 | 0.008 |
| | | (1.33) | |
| Questions Asked in Class | 21.33 | -0.69 | 0.009 |
| | | (0.76) | |
| Attendance Rate | 89.95 | 1.24 | 0.015 |
| | | (0.97) | |
| Academic Interest Index | 85.62 | -0.17 | 0.014 |
| | | (0.86) | |



*Notes:* Each row reports the coefficient on the treatment indicator from a separate regression. All regressions control for baseline values (if available), strata fixed effects, demographic controls, and school characteristics. Standard errors are clustered at the school level.

Statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

## A. Appendix

### Appendix: What Students Can Innovate

This appendix showcases a selection of student-led innovations developed during the two-year school-based innovation training program. These projects illustrate that when students are given structured opportunities for exploration, feedback, and hands-on design, they can generate original and valuable ideas—even in highly resource-constrained educational environments. The selected examples demonstrate problem identification, user empathy, analogical reasoning, and creative recombination. Projects span diverse domains, from financial inclusion and public infrastructure to mobility, hygiene, and environmental sustainability. Together, they offer a behavioral proof of concept: *innovation skills are malleable, and students can meaningfully engage in idea generation when given the right context.*

**Theme: 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. |

**Theme: 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. |

**Theme: 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. |

**Theme: 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. |

**Theme: 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. |

These projects illustrate what students can innovate when given opportunities to observe, explore, and build. They reflect a wide range of ideas grounded in the students' lived environments and community needs. The consistent use of analogical reasoning,

applied problem-solving, and user-centered thinking reinforces the central finding of this study: *innovation-related skills are teachable, measurable, and meaningfully expressed by students in low-resource public education systems.*

### A.1 Program Implementation and Monitoring

The innovation training program was implemented over two academic years, with each year comprising two thematic units and 20 sessions (40 total), each lasting 90 minutes. In Year 1 (2022–23), students developed foundational skills in problem identification and investigation through the units *Planet* and *People*. In Year 2 (2023–24), they extended these skills through *Design Evolutions* and *Smarter Everyday*, with a focus on redesign and technology integration. Each unit spanned 8–10 sessions and followed a consistent five-step design thinking cycle: FIND, LEARN, IDEATE, MAKE & TEST, and PRESENT.

Across the two years, the curricular arc was designed to reflect increasing student autonomy. Year 1 focused on structured challenges and scaffolded exercises; students were expected to develop a keen eye for identifying problems, use structured tools to investigate causes, and become confident with step-by-step problem solving. In Year 2, students worked more independently, applying design tools and integrating sensors into their solutions. They were expected to take ownership of their process and inspire others by publicly presenting their innovations.

Each unit followed a ten-session structure mapped to the FLIMP cycle. Activities progressed from guided observation and secondary research to brainstorming, prototyping, peer testing, and presentation. The table below summarizes the consistent sequence applied across all units:

TABLE A6: STANDARD 10-SESSION STRUCTURE ALIGNED WITH THE FLIMP CYCLE

| Session # | Stage | Learning Focus |
|---|---|---|
| 1 | FIND 01 | Understand the theme, identify broad challenges |
| 2 | FIND 02 | Narrow problem space using structured tools (e.g., rubric) |
| 3 | FIND 03 | Investigate causes/effects, document user perspectives |
| 4 | LEARN | Study real-world case studies, gather design inspiration |
| 5 | IDEATE | Brainstorm solutions, evaluate and select best one |
| 6 | MAKE & TEST | Build initial prototype, test with peers or users |
| 7 | PRESENT 01 | Present to peers, incorporate feedback |
| 8 | CHALLENGE | Quick redesign or open challenge (e.g., user-object pair) |
| 9 | MAKE 2.0 | Rebuild/refine based on feedback or new constraints |
| 10 | SHOWCASE | Reflect + present solution in a public forum (e.g., poster) |

Each student team worked from a structured workbook that replaced formal instruction. These workbooks contained session-level guidance, cut-and-paste tools, team reflection sheets, individual participation trackers, peer feedback rubrics, and presentation planning templates. A sample reflection prompt read: *"What one thing did your team do better today than in the last session?"*, while team rating prompts included: *"Did your team complete the task on time?"* and *"What do you want to improve next session?"* Each classroom also received an innovation kit valued at Rs. 15,000 ( $180 USD), containing circuits, sensors (light, temperature, motion, sound), wires, breadboards, and basic prototyping materials.

In Block 3, students explored the concept of redesign by analyzing how products like pens, televisions, and earphones had evolved over time. Clues such as *"There was no place to store ink inside the pen"* supported drawing-based activities that revealed how designs

improve through iteration. Students then identified real-world challenges (e.g., chalk dust, difficult-to-open containers) through structured observation and interviews, brainstormed solutions, and created mock-up or sketch prototypes. Ideas were evaluated using scoring rubrics, and resource planning sheets helped students document their material needs.

In Block 4, the focus shifted to sensor-based innovation. Students mapped each human sense to a sensor and assembled circuits using real components—buzzer, LED, motors—paired with sensors like sound, distance, light, and heat. A structured electronics manual guided this exploration. Prompts included: *"Clap near the sound sensor—what happens?"* and *"Cover the light sensor and turn the knob—how does it behave?"* Students were also introduced to storyboarding as a new prototyping mode, used to communicate their ideas through sequential diagrams.

Workbook activities like "*Smart or Not?*" encouraged critical thinking: students evaluated fictional design ideas (e.g., a buzzer that activates when books fall) to determine whether they were meaningful or silly. Students also planned and delivered presentations using scripted cue cards for each stage: problem definition, ideation, prototype description, and team roles. Final team ratings and end-of-block reflections (e.g., "What could have been better?") reinforced metacognitive awareness and helped students track their own growth.

Each classroom was supported by two student leaders (SLs) selected by teachers. SLs received 20 hours of training and participated in debriefing circles after each curriculum block. They were responsible for timekeeping, materials coordination, and non-instructional facilitation. The SL handbook emphasized six "common elements" for every session—seating, material readiness, clear opening, portfolio completion, and closing routines. SLs were instructed: *"You may clarify instructions but must not provide solutions or influence design choices."* Roles alternated between Facilitator and Observer. The latter documented classroom insights and student engagement using structured notes.

Implementation fidelity was monitored weekly through random school visits by seven trained observers using standardized rubrics. Observation protocols tracked completion of core activities, adherence to timing, material usage, and facilitation quality. In Year 1, 35.2% of sessions were observed; in Year 2, 37.5%. Across all schools, 98.1% of sessions were completed, and 88.0% of observed sessions fully implemented every prescribed element. These data, combined with workbook artifacts and session logs, confirm high adherence to the intervention design.

A stratified sample of 12 student teams was followed longitudinally, selected by baseline performance quartile. Researchers documented their progression through video, workbook analysis, and team interviews. These records provided qualitative evidence of cognitive growth, shifts in idea quality, and group-level dynamics—including collaboration patterns, time management styles, and iterative design practices.

### A.2 Qualitative Analysis of Student Feedback

To complement the quantitative findings, I collected open-ended feedback from students who participated in the innovation training program. This qualitative evidence provides additional insights into students' subjective experiences and perceived benefits of the intervention.

At endline, I asked treated students to provide written feedback about their experience with the Inqui-Lab program. This resulted in 2,524 text entries, of which 1,746 contained

substantial responses (text length $> 50$ characters). To analyze this large corpus of qualitative data systematically, I employed both computational sentiment analysis and thematic coding approaches:

1. **Sentiment Analysis:** Using GPT-4's natural language processing capabilities, I classified each substantive response as positive, neutral, or negative. This analysis revealed overwhelmingly positive sentiment, with 97% of responses coded as positive, 3% as neutral, and 0% as negative.

2. **Thematic Analysis:** Using computational assistance from GPT-4, I identified recurring themes in student responses. Five dominant themes emerged: problem-solving confidence, community responsibility, positive change, sense of possibility, and career aspirations.

Representative quotes from each theme provide deeper insight into students' experiences:

**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 problem-solving abilities (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 the application of innovation skills to address community challenges, reflecting the pattern observed in innovation content analysis 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 attributes related to self-efficacy and growth mindset.

**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 training program influenced student outcomes. The emphasis on problem-solving, community relevance, and expanded sense of possibility aligns with the theoretical framework of skill development and innovation focus discussed in the main text.

## B. Additional Outcome Measure Details

### B.1 Multi-Armed Bandit Task Implementation

The multi-armed bandit task was implemented as a business simulation with three distinct ventures: Fresh Juice Stand, Ice Cream Stand, and Pani Puri Stand. This implementation draws on March (1991)'s framework that characterizes innovation as requiring a balance between exploitation (refining existing ideas) and exploration (discovering new possibilities).

Following Ederer and Manso (2013), the task was designed so that different business choices represented different risk-reward profiles. The Fresh Juice Stand represented an exploitation strategy (lower risk, lower potential reward), while the Ice Cream Stand and Pani Puri Stand represented exploration strategies with varying degrees of risk and potential reward.

To help students understand the structure of the task, the interface provided contextualized instructions in both English and Telugu. The in-task language framed the experience as a competitive business challenge in which the goal was to earn the highest hourly profit in the class. Figure A1 shows a winner receiving a symbolic prize. Below is a translated excerpt from the instructional screen:

> **BUSINESS GAME: Maximizing Hourly Profits Challenge**
> *Imagine you're starting your very own business and want to make the most money possible. You're thinking of three types of stands: Fresh Juice, Pani Puri, and Ice Cream. Meet Prasuna. Prasuna tried the following: Type of Stand: Fresh Juice; Location: At Market; Menu Items: 5; Working Hours: 9;*

*Selling Price: Rs. 19. Prasuna's Hourly Profits: Rs. 30. Your turn: try different combinations of these inputs across 12 rounds to beat your classmates. If you achieve the highest profit, you win the first prize in your class. Good luck!*

This contextual framing reinforced the goal structure of the task while allowing students to operate in a familiar, low-stakes simulated environment.

Table A7: Probability Distribution of Profits by Business Type

| Profit (₹) | Probability (%) | | |
|---|---|---|---|
| | Fresh Juice Stand | Ice Cream Stand | Pani Puri Stand |
| 0 | 25.73 | 63.33 | 54.00 |
| 10 | 9.77 | – | 9.60 |
| 20 | 10.90 | – | 11.50 |
| 30 | 11.40 | – | – |
| 40 | 11.10 | 8.53 | 10.60 |
| 50 | 9.97 | – | – |
| 60 | 8.17 | – | 8.93 |
| 70 | 6.00 | – | 7.00 |
| 80 | 3.83 | – | 6.77 |
| 90 | 2.03 | – | – |
| 100 | 0.83 | 5.20 | 4.47 |
| 110 | 0.23 | – | – |
| 120 | 0.03 | 2.43 | – |
| 130 | – | – | 3.40 |
| 140 | – | 1.00 | – |
| 160 | – | 1.87 | 0.27 |
| 180 | – | – | 0.03 |
| 190 | – | 0.80 | – |
| 220 | – | 0.23 | – |
| 250 | – | 0.03 | – |
| Expected Value | 43.27 | 32.73 | 37.77 |
| Std. Deviation | 25.65 | 52.40 | 40.83 |

*Notes:* This table shows the fixed profit distributions used for each stand type in the bandit task. Students were not shown these probabilities but learned through feedback. The Fresh Juice Stand offered safer, lower-variance returns; the Ice Cream and Pani Puri Stands featured higher variance and higher maximum potential profits, with frequent zero-profit outcomes. These distributions were pre-programmed and identical across students.

Each round, students selected a stand type, location, working hours, menu size, and price. These variables jointly influenced the probability of receiving specific profit values based on the underlying distribution of the selected stand type. Students were not informed of the exact functional form, but learned through repeated feedback over 12 rounds. Table A8 provides an example sequence of decisions and outcomes.

Table A8: Example Decision Sequence in Multi-Armed Bandit Task

| Round | Type of Stand | Location | Working Hours | Menu Items | Price | Profit |
|---|---|---|---|---|---|---|
| 1 | Pani Puri | School | 8hrs | 12 | ₹13 | ₹20 |
| 2 | Ice Cream | Market | 12hrs | 7 | ₹16 | ₹70 |
| 3 | Fresh Juice | Homes | 9hrs | 6 | ₹13 | ₹40 |
| 4 | Ice Cream | School | 10hrs | 11 | ₹16 | ₹0 |
| 5 | Ice Cream | School | 11hrs | 5 | ₹13 | ₹40 |
| 6–12 | Subsequent rounds... | | | | | |

*Notes:* This example illustrates how students interact with the task across multiple rounds. Each round involves choosing a business type, location, hours, menu size, and price. Students receive immediate profit feedback after each choice and can view their history to guide future decisions. Prior selections are pre-filled in subsequent rounds, though students can adjust all inputs freely.

Students saw immediate feedback after each round, including the profit earned and a full history of previous choices and profits. Inputs were pre-filled based on their previous round, though students could modify any variable. The task interface was built using JavaScript and deployed offline via Qualtrics on tablets. Prior to the experiment, I piloted the task with 35 students from similar backgrounds to confirm understanding, variation in behavior, and usability of the interface.

To simulate real stakes, the top-performing student in each class received a small prize. A photograph of one winner receiving this prize is shown below to illustrate the incentive mechanism.



Figure A1: Student winner receiving prize for highest profit in the bandit task

### B.2 Innovation-S Scale Development Process

The development of the Innovation-S scale followed a rigorous multi-stage process. After conducting in-depth interviews with 14 successful innovators, I identified key themes and potential scale items. These were then refined through cognitive interviews to ensure clarity and relevance. The final survey was administered to 44 experienced innovators representing diverse professional backgrounds. The industry distribution included education (40.9%), technology (15.9%), healthcare (9.1%), finance (6.8%), manufacturing (4.5%), and others (22.8%). Participants held various roles including founders/CEOs (25.0%), product managers (13.6%), analysts/researchers (11.4%), educators (20.5%), and engineers (9.1%).

The initial pool contained 15 potential scale items, which factor analysis reduced to a 7-item scale with a two-factor structure. Factor 1, labeled "Novelty," included items

such as "Offers a new solution to a problem," "Applies existing knowledge in novel ways," "Challenges conventional approaches," and "Introduces a new perspective or approach." Factor 2, labeled "Value-Add," included items such as "Improves the lives of those who will use it," "Addresses a clear need or pain point," and "Produces meaningful, measurable benefits."

The scale demonstrated acceptable internal consistency (Cronbach's alpha = 0.69) and good construct validity. Principal component analysis yielded a Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy of 0.57 and Bartlett's test of sphericity was significant ($p < 0.001$), supporting the factor structure. Validation against real-world outcomes showed strong correlations with reported user base ($r = 0.71$, $p < 0.001$) and funding raised ($r = 0.31$, $p < 0.05$), indicating good predictive validity. These associations held across innovations of varying scales, from those serving a few dozen users to those reaching millions, and funding levels ranging from a few thousand to several million dollars.

### B.3 Team Idea Data Collection, Processing, and Evaluation Process

Each student team submitted their innovation idea using a standardized format that included an audio recording and a hand-drawn sketch or diagram. The audio was structured around four guiding prompts: identifying the user, articulating the problem, proposing a solution, and explaining the innovation's value. Students recorded their explanations in their native language, following a script provided in advance to ensure consistency across submissions. Most teams submitted one to two pages of drawings or schematics alongside the audio (see Figure A2). These images provided a visual representation of their ideas and supported evaluators in understanding prototypes, even when physical models were not developed.

The audio files were transcribed and translated by bilingual research assistants using a standardized protocol. Transcriptions preserved the students' wording while ensuring linguistic clarity and cross-submission comparability. Audio and visual materials were matched and anonymized prior to evaluation. Quality checks included spot audits and standardization reviews by senior research staff.

Each idea was independently evaluated by two trained raters who were blind to treatment status. Raters scored submissions using two instruments: the Innovation-S scale—a psychometric tool developed through expert calibration—and a simulated VC funding allocation exercise, in which raters distributed up to 1000 per idea based on perceived impact, novelty, and feasibility. Although scores were continuous, raters followed a qualitative rubric with categories such as "Exceptional," "Good," "Feasible," "Vague," and "Risky," to guide allocations. Funding scores did not always align with creativity alone; raters also considered clarity of presentation, implementation cost, and prototype feasibility.

Rater memos indicate recurring themes across ideas, including social inclusion, women's safety, water conservation, farming innovation, and environmental protection. Many students proposed sensor-based solutions to familiar problems, often reflecting the challenges of their rural and agricultural contexts. Even where ideas were basic or unpolished, raters noted that student intent and contextual relevance shaped funding decisions. This multi-stage process—structured submission, standardized processing, and double-blind evaluation—provided a credible and

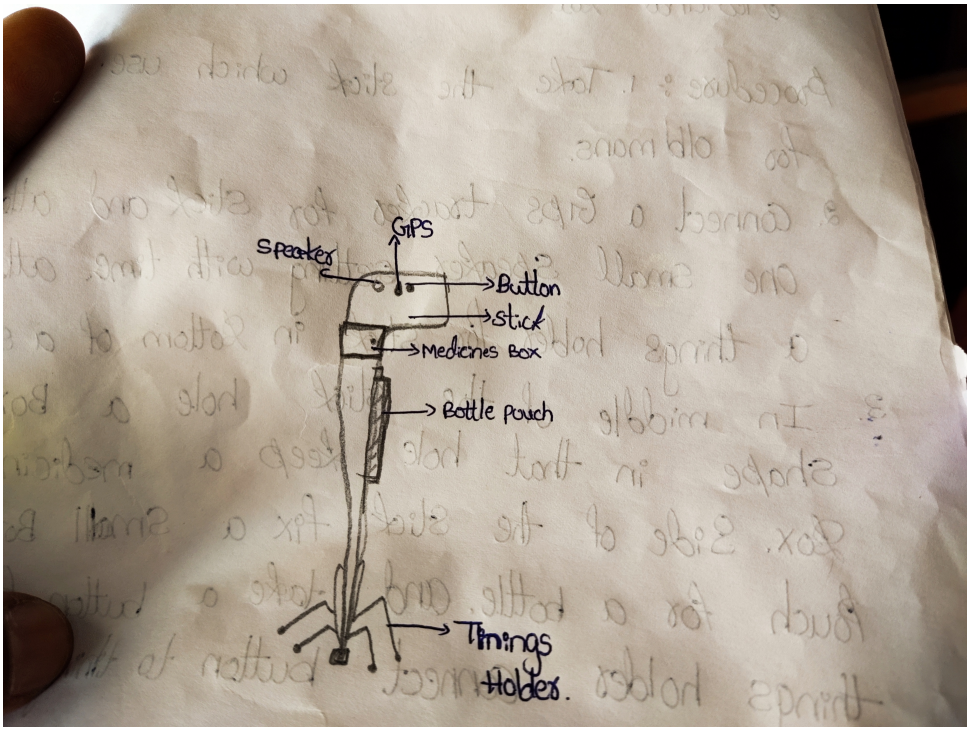



Figure A2: Smart Medication Reminder Walking Stick. An assistive device combining a traditional walking stick with medication storage compartment, audible reminder system, and GPS tracking functionality to help elderly individuals maintain medication adherence while living independently.

### B.4 Fundraising Challenge Implementation Details

The inter-school fundraising challenge was designed to simulate real-world funding processes for innovations, providing an externally valid measure of innovation quality beyond laboratory assessments. From the 1,200 innovation ideas collected across treatment (T) and control (C) groups, I selected the top 40 ideas from each condition based on their Innovation-S scale ratings by independent evaluators. This selection process ensured that only the highest quality ideas from each group competed in the challenge, while maintaining balanced representation between treatment and control conditions.

To minimize preparation bias and simulate realistic innovation pitching conditions, teams received notification only 5 days before the event. This short notice period prevented differential preparation advantages and better reflected real-world innovation competitions where teams must quickly prepare presentations from existing work. Team size was standardized to 3-4 students per team, with members selected from the original innovation teams that developed each idea. All teams received identical presentation templates and guidelines: 5-minute pitch time, standardized slide format (problem statement, solution description, prototype details, potential impact), and 3-minute Q&A session with judges. Teams presented in randomized order, with treatment and control teams interspersed to prevent order effects. Judges were blind to treatment condition. Judges used a standardized scoring rubric using the innovation scale and funding they want to grant. Each dimension was rated on a 1-5 scale. Each judging pair had a fixed budget of  20,000 to allocate across 20 teams, with no minimum or maximum per team, requiring judges to make relative value judgments similar to real investment decisions.

The expert jury comprised eight professionals with extensive experience evaluating

early-stage innovations. This included three non-profit entrepreneurs who were founders or senior leaders at established social enterprises with 5-10 years of experience implementing innovation-based solutions in education, healthcare, and rural development. Two social impact investors participated as investment managers from social impact funds with experience evaluating early-stage ventures in the 5-50 lakh (approximately $6,000-$60,000) funding range. Additionally, three corporate innovation leaders served as jurors, representing directors or managers of innovation departments at major Indian companies, responsible for evaluating internal and external innovation proposals. Jurors were recruited through professional networks and selected based on criteria including: (1) minimum 5 years of experience evaluating innovations, (2) familiarity with adolescent innovation competitions or education initiatives, and (3) no prior relationship with the research project or implementing partners to ensure impartiality.

I implemented a double-blind assessment protocol to minimize potential biases. Judges were not informed about the research design or which teams belonged to treatment vs. control conditions. Each pair of jurors evaluated exactly 20 teams, with teams balanced across treatment and control groups (10 from each condition). Judges were randomly assigned to team groupings, and the assignment of judges to teams was stratified to ensure balanced expertise across innovation domains. Judges independently scored each presentation before conferring on funding allocations, providing both individual assessments and collective funding decisions.

The primary outcome measure from this challenge was the treatment-control difference in average fundraising amounts, representing the real-world valuation gap between innovations from treatment and control students. Secondary measures included average scores on individual assessment dimensions and qualitative feedback from judges on innovation quality. This multi-faceted approach to measuring innovation quality through expert assessment and funding allocation provided externally valid evidence on the effectiveness of the innovation training program, complementing the more controlled laboratory and survey-based measures.

### B.5 Measurement Adaptation Procedures

The original RMET, developed by Baron-Cohen et al. (2001), contains 36 black and white photographs of the eye region of Caucasian adults, with each image accompanied by four emotion words. To ensure cultural relevance and age appropriateness for Indian adolescents, I created an adapted version that retained the 12 highest-discrimination items from the original test and added 8 new items featuring South Asian faces. This adaptation process included ensuring age-appropriate vocabulary for the emotion words and validating translation of emotion terms with bilingual experts. The cultural adaptation approach follows best practices established by Vellante et al. (2013) for cross-cultural application of emotional intelligence measures. The adapted version was piloted with 148 out-of-sample students from the target population, achieving a Cronbach's alpha of 0.71, comparable to the original test's reliability in adolescent populations. Item-total correlations ranged from 0.23 to 0.58, indicating good discrimination.

I used selected items from Raven's Advanced Progressive Matrices (Set II) to create a 20-item assessment appropriate for the target population. Items were selected based on increasing difficulty to minimize ceiling effects, coverage of different pattern types (e.g., addition/subtraction, figure progression), and results from pilot testing with 50 students

Table A9: Outcomes Landscape: Measures and Their Sources

| **Category** | **Measure** | **Source/Methodology** |
|---|---|---|
| Innovative | Multi-Armed Bandit Task | March (1991) and Ederer and Manso (2013) |
| Ability | Quality of Ideas Scale | Novel scale developed with input from social entrepreneurs, scientists, and patent authors |
| | Interschool Fundraising | Expert jury assessment |
| Higher-Order | Problem Solving | Sudoku puzzles (Newell and Simon, 1972; Yato and Seta, 2003) |
| Skills | Fluid Intelligence | Raven's Matrices (Raven, 2000) |
| | Emotional Intelligence | RMET (Baron-Cohen et al., 2001) |
| | Grit | Behavioral measure (Alan et al., 2019) |
| Academic | Mathematics | Adapted from Muralidharan et al. (2019) |
| Outcomes | Science | Curriculum-based assessment (State Council of Educational Research and Training, Telangana, 2020) |
| Self-Reported | Attendance | Self-reported data |
| Measures | Classroom Engagement | Self-reported "questions asked to teachers" |
| | Career Aspirations | Self-reported future goals |

*Notes:* This table summarizes the outcome measures used in this study across four categories. The Innovative Ability measures directly assess students' capacity to explore, create, and develop novel solutions. Higher-Order Skills measure cognitive and socio-emotional abilities related to innovation. Academic Outcomes assess potential spillover effects to traditional subjects. Self-Reported Measures capture student engagement and aspirations.

from similar schools. The adapted version achieved a Cronbach's alpha of 0.82 in pilot testing, with good discrimination across the ability spectrum of our target population.

The science assessment was developed through a systematic process that included review of state curriculum and textbooks for grades 8-10, identification of key concepts across physics, chemistry, and biology, development of 30 multiple-choice and 5 short-answer questions, expert review by 3 science teachers from non-participating schools, and pilot testing with 65 students from similar backgrounds. The final assessment included 25 items covering fundamental scientific principles, with a Cronbach's alpha of 0.76 and appropriate difficulty distribution (mean score 58% in pilot sample).

### B.6 Midline Results and Progression Over Time

I collected midline data after the first year of the intervention to examine how treatment effects evolved over the course of the program. Comparing midline and endline results reveals an interesting progression that provides insight into skill development trajectories.

At midline, the innovation program had already produced significant improvements in students' innovation scores (2.03 percentage points, $p<0.05$) and related behaviors, including exploitative approach (4.82 percentage points, $p<0.01$) and explorative approach (2.55 percentage points, $p<0.05$). However, midline results showed a negative effect on "Started With Exploration" (-0.031, $p<0.05$), which reversed to a positive effect by endline (0.060, $p<0.01$). This suggests that students initially approached the innovation tasks conservatively but developed more sophisticated innovation strategies by the end of the two-year program.

A particularly striking contrast appears in the higher-order skills domain. At midline, the program showed no significant positive effects on fluid intelligence (-1.75 percentage points, not significant) and a negative effect on emotional intelligence (-1.72 percentage points, $p<0.05$). By endline, these patterns had reversed, with significant positive effects on fluid intelligence (2.63 percentage points, $p<0.05$) and problem-solving (3.67 percentage points, $p<0.05$). This progression suggests that cognitive benefits from the innovation training emerged only after sustained engagement with the program, consistent with research on neuroplasticity indicating that cognitive changes often require extended practice.

The academic outcome patterns also evolved meaningfully over time. At midline, the program had a negative effect on math scores (-2.82 percentage points, $p<0.01$) and math interest (-0.16, $p<0.01$). By endline, these negative effects had disappeared in the full sample, with math scores showing a small, non-significant positive effect (0.16 percentage points). This pattern suggests that initial trade-offs between innovation activities and traditional academics diminished over time, possibly as students learned to integrate these different forms of learning.

Student engagement metrics showed similar evolution. While attendance showed no significant treatment effect at midline (-0.17 days, not significant), by endline there was a marginally significant positive effect (1.78 percentage points, $p<0.10$).

These temporal patterns provide important context for interpreting the main results. The contrasting midline and endline effects suggest that some skills require sustained engagement to develop, while others may involve initial trade-offs that diminish over time. The progression also indicates that the second year of the program was crucial for developing higher-order cognitive skills and resolving potential tensions with traditional

academic performance.

Table A10: Treatment Effects on Student Outcomes in Year 1

| Category | Outcome | Control Mean | Treatment Effect | $R^2$ |
|---|---|---|---|---|
| **Panel A: Innovation Outcomes** | | | | |
| Innovation Measures | Innovation Score | 57.66 | 2.029** | 0.033 |
| | | | (0.714) [0.011] | |
| | Exploitative Approach | 31.80 | 4.820*** | 0.063 |
| | | | (1.587) [0.006] | |
| | Explorative Approach | 64.35 | 2.550** | 0.029 |
| | | | (1.126) [0.031] | |
| | Started With Exploration | 0.43 | -0.031** | 0.013 |
| | | | (0.015) [0.038] | |
| **Panel B: Higher-Order Skills** | | | | |
| Cognitive and Non-cognitive Skills | Fluid Intelligence | 54.31 | -1.748 | 0.074 |
| | | | (1.283) [0.233] | |
| | Emotional Intelligence | 44.16 | -1.721** | 0.094 |
| | | | (0.750) [0.035] | |
| | Big 5: Extraversion | 3.21 | -0.035** | 0.042 |
| | | | (0.014) [0.019] | |
| | Big 5: Agreeableness | 3.32 | -0.002 | 0.079 |
| | | | (0.017) [0.931] | |
| | Big 5: Conscientiousness | 3.23 | -0.017 | 0.058 |
| | | | (0.015) [0.278] | |
| | Big 5: Neuroticism | 2.85 | 0.035** | 0.011 |
| | | | (0.014) [0.028] | |
| | Big 5: Openness | 3.25 | 0.015 | 0.104 |
| | | | (0.020) [0.522] | |
| **Panel C: Academic Outcomes** | | | | |
| Academic Performance and Engagement | Math Score | 45.30 | -2.819*** | 0.175 |
| | | | (0.917) [0.013] | |
| | Science Score | 50.92 | 0.394 | 0.157 |
| | | | (0.662) [0.541] | |
| | Liked Math | 2.55 | -0.158*** | 0.018 |
| | | | (0.034) [0.000] | |
| | Liked Science | 2.50 | -0.044 | 0.008 |
| | | | (0.039) [0.307] | |
| **Panel D: Survey Responses** | | | | |
| Student Behavior and Aspirations | Attendance | 10.36 | -0.168 | 0.011 |
| | | | (0.149) [0.315] | |
| | Questions Asked | 3.18 | -0.186* | 0.014 |
| | | | (0.093) [0.059] | |
| | Number of Friends | 4.71 | -0.129 | 0.050 |
| | | | (0.162) [0.448] | |
| | STEM Career Aspiration | 0.33 | 0.002 | 0.197 |
| | | | (0.014) [0.913] | |
| | Teacher Career Aspiration | 0.06 | 0.011 | 0.153 |
| | | | (0.008) [0.158] | |
| **Panel E: Risk Behaviors** | | | | |
| Risk Measures | Risk Index | 0.05 | -0.092 | 0.019 |
| | | | (0.074) [0.265] | |
| | Super-betting | 3.21 | -0.109 | 0.011 |
| | | | (0.074) [0.169] | |

*Notes:* Each cell reports the coefficient on the treatment indicator from a separate regression.
Values in parentheses are standard errors, clustered at the school level.
The value in square brackets is the wild bootstrap p-value, calculated using 1000 replications.
Statistical significance is denoted by $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.
All regressions include strata fixed effects, the outcome's baseline value when available, demographic controls, and school characteristics.
Control mean reports the mean value for the outcome variable in the control group in Year 1 (midline).

## Appendix G. A Field Protocol for Tablet-Based Educational Assessment for Measuring Higher-Order Skills

This section documents the field protocol used to implement large-scale, tablet-based student assessments for measuring higher-order skills in government schools serving marginalized communities. The approach was developed and deployed as part of the baseline, midline, and endline data collection. It offers a replicable blueprint for conducting rich, technology-enabled assessments in low-resource educational settings.

All student assessments were administered using the Qualtrics Offline App, which enabled survey delivery in low-connectivity environments. Each student received an individual tablet preloaded with the complete survey in both English and Telugu. The assessment included six core modules, which were delivered in a randomized order to mitigate survey fatigue and ordering effects. The modules included: (1) a multi-armed bandit task measuring innovation strategy and value discovery, (2) a problem-solving and grit module using timed Sudoku puzzles, (3) the Raven's Progressive Matrices IQ test, (4) the Reading the Mind in the Eyes Test (RMET) for emotional intelligence, (5) a 10-item Big Five personality inventory, and (6) subject-specific quizzes in mathematics and science.

Randomizing the order of modules served a second purpose: it reduced opportunities for cheating by ensuring that students seated next to one another encountered different content at any given time. This structure also prevented leakage of questions to teachers or peers—a known limitation of traditional paper-based assessments in school settings. In addition, the interface presented one question per screen, further minimizing copying and improving response accuracy. These design choices draw on evidence from Singh et al. (2024), which demonstrates that tablet-based assessments reduce cheating and improve measurement validity relative to paper-based formats. This approach allowed for high-quality, secure data collection at scale.

Enumerators underwent a multi-day training that included protocol walkthroughs, live classroom simulations, and tablet handling sessions. The training emphasized data fidelity, neutrality, and student protection. All enumerators received detailed manuals in both English and Telugu, covering everything from survey script delivery to device troubleshooting. Field teams followed a structured appointment protocol with school principals, including scripted calls to confirm logistics and prepare students in advance. Surveys were conducted simultaneously across classrooms using charged and preconfigured tablets, with strict protocols to prevent teacher interference.

On the day of data collection, enumerators set up chart paper or used the blackboard to display general information to reduce repetition across students. All instructions were read aloud and supplemented with visual cues. Students completed the tablet survey independently but could raise their hand to ask questions. Scheduled breaks were built into the process, and high performers were awarded small prizes (e.g., pens or certificates) to sustain engagement. Enumerators collected tablets immediately after completion, checked for data integrity, and uploaded results upon return to a connected location.

In parallel with the individual assessments, enumerators also collected team-submitted innovation ideas. Students submitted an audio explanation and hand-drawn sketch following a structured format: user, problem, solution, and innovation value. Audio was recorded in the student's preferred language and later transcribed and translated by



(a) Enumerator briefing students using the blackboard



(b) Tablet-based assessments in progress



(c) Students seated with individual tablets



(d) Students completing modules independently

Figure A3: Implementation of tablet-based assessments across government schools. Each student worked on an individual device in randomized module order using the Qualtrics offline app. Enumerators explained instructions on a blackboard or chart paper before starting.

bilingual research assistants. The visual submission was captured in landscape mode and included only the drawing, idea title, team member names, school, and class section. Raters evaluated these ideas blind to treatment status using a psychometric rubric (Innovation-S scale) and a simulated VC funding exercise.

Teams implemented a split-shift model, dividing responsibilities between tablet as-

sessments and idea collection. Each enumerator was responsible for tablet condition, data saving, and respondent behavior. Daily debriefing calls with field coordinators allowed rapid feedback on data quality, technical issues, and scheduling. These protocols ensured high-fidelity implementation across diverse school settings, even with limited infrastructure.

This protocol demonstrates the feasibility of capturing rich, behaviorally grounded measures of student cognition and creativity using tablet-based tools in public schools without WiFi. It also shows how to integrate team-based innovation data with randomized assessments at scale. The design choices outlined here reflect seven years of field experience implementing technology-enabled assessments in Indian government schools. In prior work, I served as the program lead for the evaluation of tablet-based assessment methods documented in Singh et al. (2024), and this study extends those principles to the measurement of higher-order skills and innovation capacity. The structure outlined here may be useful for researchers conducting field experiments involving low-stakes cognitive or behavioral tasks, especially in contexts where conventional infrastructure is limited and linguistic diversity is high.

**Appendix: Cost-Effectiveness and Illustrative Returns**

**Program Cost and Per-Student Investment.** The innovation program cost approximately $40 per student annually, primarily covering classroom materials, peer facilitator training, and monitoring. As implemented over two academic years, the total per-student cost was $80. These figures are based on detailed budget records from Inqui-Lab Foundation during the intervention period.

**Cost per Standard Deviation Gain.** Across the three core innovation outcomes, the average standardized treatment effect was approximately 0.22 SD. This implies a cost-effectiveness ratio of $364 per standard deviation gain ($80 ÷ 0.22). Expressed in conventional education impact terms, this corresponds to roughly $36 per 0.1 SD improvement in innovation-related skills. These estimates are not directly comparable to learning-adjusted-years-of-schooling (LAYS) metrics, but fall within the range of highly cost-effective education programs benchmarked in Evans and Yuan (2022).

**Comparison with Other Interventions.** Computer-assisted learning programs evaluated in similar Indian contexts cost between $100–$200 per student and yielded learning gains of 0.1–0.3 SDs (Muralidharan et al., 2019). Remedial instruction programs cost around $67 per student and generated test score gains of 0.14–0.28 SDs (Banerjee et al., 2007). While these programs target academic domains, the innovation program compares favorably in terms of effect size and unit cost for non-traditional skills.

**Statewide Implementation Cost.** Scaling the program to all Grade 8 students in Telangana (approximately 400,000 students) would cost an estimated $16 million annually. This represents approximately 0.6% of the state's education budget and is comparable to other government-funded innovation programs, including the Atal Tinkering Labs (national budget of $240 million) and Delhi's Entrepreneurship Mindset Curriculum ($27 million annually).

**Illustrative Economic Returns (Highly Speculative).** While the study does not estimate long-run earnings impacts, a back-of-the-envelope calculation illustrates potential returns under conservative assumptions. Suppose 20% of participants experience a 5–10% increase in productivity in adulthood, consistent with documented gains from soft skills and managerial training in small enterprise settings (Bloom et al., 2013; Jensen and Miller, 2018; Adhvaryu et al., 2018; Atkin et al., 2017). If these gains are realized among informal sector workers (88% of the labor force (ILO, 2018)) earning $2,000 annually, this would represent $100–$200 in annual productivity gains for a subset of workers. Aggregated over a working life, this would plausibly offset the cost of the intervention many times over.

These estimates are purely illustrative. They assume persistence of skill gains and successful transition into relevant labor market roles, neither of which are directly tested in this study. They are included to demonstrate plausible economic relevance, not to claim measured long-term impacts.

**Power Calculation Details**

I calculated the minimum detectable effect size (MDES) using the formula for cluster-randomized trials as described in Dong and Maynard (2013). This approach incorporates key design parameters including the desired significance level, desired power, intraclass correlation coefficient (ICC), proportion of variance explained by covariates, average cluster size, treatment allocation ratio, and number of clusters.

The following inputs were used based on the study design and implementation context:
- 80 schools (40 treatment, 40 control),
- average cluster size of 55 students based on actual enrollment,
- ICC of 0.10 (a plausible value for cognitive outcomes in educational studies; Hedges and Hedberg 2007),
- $R^2 = 0.50$ based on the inclusion of strong baseline outcome controls,
- equal allocation to treatment and control (50% each),
- desired power of 80%, and
- significance level of 5% (two-tailed).

The MDES was computed using the formula:

$$\text{MDES} = (t_{\alpha/2} + t_{\beta}) \cdot \sqrt{\frac{(1-R^2)(1+(m-1)\cdot \text{ICC})}{P(1-P)\cdot J \cdot m}}$$

$$= (1.96+0.84)\cdot\sqrt{\frac{(1-0.50)(1+(55-1)\cdot 0.10)}{0.25\cdot 80\cdot 55}} = 2.80\cdot\sqrt{\frac{3.2}{1100}} = 2.80\cdot 0.05396 = \boxed{0.151}$$

This result implies that the study had 80% power to detect an effect size of 0.151 standard deviations at the 5% significance level. This level of sensitivity is sufficient to detect the magnitude of effects observed for several of the program's key outcomes (e.g., innovation scores, fluid intelligence, and problem-solving).

This MDES is also within the range of effects typically considered educationally meaningful (0.15–0.20 SD; Hill et al. 2008). Moreover, the inclusion of pre-specified baseline outcomes, use of stratified randomization, and high implementation fidelity may have further increased statistical power. Nonetheless, actual power may vary depending on the realized ICC and sample composition at endline.

Table A11: Minimum Detectable Effect Size (MDES) Under Varying Assumptions

| $R^2$ | **ICC = 0.05** | **ICC = 0.10** | **ICC = 0.15** |
|---|---|---|---|
| 0.30 | 0.136 | 0.179 | 0.213 |
| 0.40 | 0.126 | 0.165 | 0.197 |
| 0.50 | 0.115 | 0.151 | 0.180 |
| 0.60 | 0.103 | 0.135 | 0.161 |