Enhancing the Quality of Thinking:
Evidence from a School-Based Cognitive Expansion Program *

Sule Alan[†] and Merve Derebasoglu[‡]

Abstract

Long-run human progress has depended on our higher-order skills, including abstract reasoning, categorizing information, drawing analogies across settings, and critically evaluating evidence. Although higher-order skills are fundamental to creativity and innovation, formal education places little direct emphasis on facilitating their development. We evaluate an educational intervention that systematically exposes students and teachers to sustained and progressively more demanding cognitive activities over two consecutive years. The intervention is designed to strengthen children's capacity for higher-order reasoning alongside the regular curriculum. The study follows more than 14,500 students for three years in Turkey. It covers 132 primary schools, randomly assigning 67 schools to the program. We estimate the impacts on a broad set of cognitive, academic, and higher-order thinking outcomes. Central to our measurement is a novel method for assessing critical thinking in children. Our language-based tool assesses logical inference, argument justification, and the detection of flawed reasoning. We find large gains in higher-order reasoning, with effects of 0.16–0.28 standard deviations on abstract and analogical reasoning. We also find improvements in working memory and attention, as well as significant gains in academic learning and critical thinking. The results suggest that sustained exposure to high-quality, cognitively demanding educational experiences can improve children's reasoning skills and academic learning.

JEL Codes: I21, J24, C93

Keywords: critical thinking, higher-order reasoning, learning

---

*We thank the European Research Council for funding the study. The project has ethics approvals from the European University Institute, Italy and Bilkent University, Turkey. We are grateful to Yusuf Agus, Neslihan Gurler, Dilasa Karakaya, Buse Tanca and numerous field assistants for their efforts and care in onerous data collection. The trial has been registered at the AEA Registry: AEARCTR-0009140.

[†]Corresponding Author: Cornell University and NBER, Email: sulealan.academic@gmail.com.

[‡]Goettingen University.

1 Introduction

Humans have a unique capacity for complex thought. We can abstract, draw analogies across settings, categorize information, evaluate evidence, and detect flaws in reasoning. These *higher-order thinking skills* allow us to solve novel problems and generate new knowledge. They are fundamental to long-run human progress. Their development is supported by the neuroplasticity of the brain, which is strongest during childhood but persists well beyond it (Knudsen, 2004; Hensch, 2005; Heckman, 2006; Lövdén et al., 2010; Fuchs and Flügge, 2014). Formal education can accelerate this process through cognitively demanding learning experiences, such as exposure to rich language, complex problem-solving and collaborative learning (Flynn, 1987; Neisser, 1998). These skills may be particularly valuable in an environment where access to information is increasingly costless. When knowledge is readily available, the ability to evaluate information, to distinguish evidence from assertion, and to detect flawed reasoning becomes increasingly important. Modern education systems tend to treat higher-order thinking skills as a natural product of academic learning rather than skills to be developed directly. As a result, education continues to focus primarily on knowledge acquisition and test scores, with little emphasis on deliberately strengthening the reasoning processes that support learning.

This paper explores whether schools can directly target and improve the quality of children's reasoning. We evaluate "Developing Brain," a program designed to push the cognitive limits of both students and teachers over two consecutive years. The study took place in two provinces in Turkey and covered 132 primary schools, more than 14,500 students, and over 550 teachers. Schools were randomly assigned to treatment and control groups. Teachers in treated schools received intensive training before grade 2 and again before grade 3 and implemented the program for at least one class hour each day. In grade 2, the program targets pattern recognition and analytical categorization, with some exercises also placing demands on memory. In grade 3, it moves to more complex forms of reasoning, particularly logical and analogical reasoning. The activities are almost entirely nonverbal and visual: children work with shapes, pictures, patterns, cards, tables, and other visual materials rather than reading or doing numerical exercises. They identify increasingly complex patterns, classify objects according to different dimensions, detect inconsistencies in visual information, and identify relationships that can be transferred across settings.

The activities become progressively more challenging for both students and teachers.

This progression in difficulty is a defining feature of the intervention. Teachers were trained to maintain an appropriate level of cognitive challenge: activities that were too easy would not expand cognitive capacity, whereas activities that were too difficult would discourage learning. They received training on the neural mechanisms of learning and brain plasticity and worked through the exercises themselves before using them in their classrooms. The program therefore does not directly target literacy or numeracy and provides no remedial content. Its overarching goal is to expand children's capacity to reason so that they become better learners. Control teachers continued with business-as-usual instruction.

The intervention displaced some curricular instruction in treated schools, particularly in grade 3. This tradeoff was deliberate. Our hypothesis was that strengthening children's reasoning capacity would allow them to learn the official curriculum more efficiently. If successful, the program would improve both learning and higher-order reasoning despite reducing instructional time devoted to the standard curriculum.[1]

We evaluate the program using four waves of data collected over three years. We measure executive functions (attention and working memory), academic performance, and a broad set of higher-order thinking skills. As children grow older, we progressively expand our measurement system to capture more sophisticated reasoning abilities. We measure abstract reasoning (fluid intelligence) and cognitive empathy (theory of mind) at every wave, including baseline. At the end of grade 2, we additionally measure attention, working memory, and analogical reasoning. At the end of grade 3, we further expand our toolkit to include critical thinking. Finally, at the end of grade 4, we enrich our critical thinking task and evaluate the persistence of the program one year after the intervention ended. Academic performance is measured in every wave, including baseline, with tests aligned with the national curriculum. In grade 2, we focus on foundational literacy and numeracy. In grades 3 and 4, we introduce increasingly demanding verbal and math tests, including questions requiring transitivity and logical reasoning. These later assessments also require language proficiency. These questions aim to capture not only mastery of the curriculum but also the development of more advanced reasoning skills.

Central to our evaluation is to assess the development of critical thinking. Despite its

[1]One motivation for the program came from repeated complaints by primary school teachers that it has become increasingly difficult to teach the same material to new cohorts of students. This concern emerged frequently in our focus group discussions with teachers. These discussions strengthened our case to the Ministry of Education to pilot a program that deliberately increases cognitive demands on students and teachers. The idea was that strengthening higher-order reasoning could accelerate academic learning.

central role in learning and decision making, no established instruments exist for measuring critical thinking in young children. We therefore develop a novel language-based assessment designed specifically for our age group. The assessment combines two complementary tasks that together capture three core components of critical thinking: (i) evaluating evidence and drawing logical inferences, (ii) justifying one's reasoning, and (iii) detecting flawed reasoning. The first task is an incentivized mystery-solving exercise in which children sequentially evaluate evidence, eliminate inconsistent hypotheses, and justify their decisions. The second task requires children to identify subtle logical inconsistencies embedded in short narratives. Both tasks require children to reason through novel situations rather than recall previously learned knowledge. We combine performance across the tasks into a single critical thinking index. We validate the critical thinking index using our rich set of cognitive, academic, and socio-cognitive measures. The index is strongly associated with executive functions (attention and memory), academic achievement, abstract and analogical reasoning, and cognitive empathy, supporting the validity of our measure.

We find that the program significantly improves higher-order thinking skills. Averaged across the three follow-up waves, treated children score 0.24 sd higher in abstract reasoning (fluid IQ), 0.21 sd higher in analogical reasoning and 0.09 sd in cognitive empathy. The gains in abstract reasoning emerge during the intervention, reaching 0.28 sd at the end of grade 2 and 0.22 sd at the end of grade 3, and persist at 0.16 sd at the end of grade 4, one year after the program ended. We find a similar pattern for analogical reasoning, with significant gains at every follow-up wave (0.18 sd, 0.17 sd and 0.16 sd, respectively). Thus, improvements in higher-order reasoning emerge during the intervention and remain significant one year after its completion, at the end of primary school.

The treatment effects also extend to other cognitive skills and academic learning. Working memory and attention, the two executive functions that underpin learning, improved immediately after the first year of the program.[2] Averaged across the three follow-up waves, treated children score 0.09 sd and 0.04 sd higher than control children in working memory and attention tests, respectively. The effects on working memory and attention dissipated by the end of primary school. Moreover, despite some crowding out of regular teaching, treated students exhibit significant gains in academic learning (0.09 sd in literacy and 0.08 in numeracy on average). Finally, at the end of primary school, we find that treated students develop stronger critical thinking skills. The critical thinking gains span all measured components:

[2]These executive functions are widely regarded as core cognitive foundations of learning.

evidence evaluation and logical inference (0.09 sd), argument justification (0.06 sd), and detection of flawed reasoning (0.06 sd). These downstream gains follow a striking pattern by language environment. Among children learning in a second-language environment (L2), we find no significant gains in either academic achievement or critical thinking, despite substantial improvements in non-language-based higher-order reasoning. This pattern suggests that cognitive expansion occurs in both language environments, but that limited proficiency in the language of instruction may limit how improvements in nonverbal reasoning translate into language-dependent skills.

The effects and their persistence may arise through multiple, potentially reinforcing channels. One possibility is that the intervention made teachers more effective by strengthening their reasoning skills and shifting their beliefs about the learning process. A second possibility is that cognitive expansion generates dynamic complementarities within children. Improvements in reasoning may accelerate learning, while learning itself may further strengthen cognitive and higher-order reasoning skills. These channels are not mutually exclusive: improvements in teachers may strengthen the dynamic process of skill formation within children. Consistent with the teacher channel, treated teachers perform better in abstract reasoning, analogical reasoning, and cognitive empathy, with effects of 0.08, 0.20, and 0.18 sd, respectively. The effect on abstract reasoning is not statistically significant. We also find a 0.26 sd increase in teachers' growth mindset, reflecting a stronger belief that intelligence can be expanded through effort.

To study the dynamic process of skill formation within children, we estimate a model of human capital accumulation, following Cunha and Heckman (2007, 2008); Cunha et al. (2010); Heckman et al. (2013); Attanasio et al. (2020); Agostinelli and Wiswall (2025). The model is motivated by the structure of the intervention. The central hypothesis is that challenging children's nonverbal reasoning skills strengthens these skills over time and also improves academic learning. Academic learning may then further expand higher-order reasoning. This creates a dynamic process in which learning and reasoning reinforce one another over time, with critical thinking emerging as children develop stronger language skills. We estimate these relationships in a dynamic framework and show that improvements in higher-order skills persist over time and contribute to later improvements in academic skills and critical thinking.

Our paper makes three contributions to the education and human capital literature. First, we introduce and evaluate an approach to improving reasoning skills. Socioeconomic

disadvantage is associated with lower levels of cognitive stimulation and weaker development of cognitive and socio-cognitive skills (Flood et al., 2022; Falk et al., 2021). Our program exposes socioeconomically disadvantaged children to sustained and progressively more demanding cognitive activities designed to strengthen abstraction, analogical reasoning, and logical reasoning. While cognitive science has long documented strong associations between these skills and academic success, evidence on whether they can be deliberately developed at scale in the classroom remains limited.[3] Using a large-scale randomized trial, we show that higher-order reasoning skills are malleable and can be strengthened through sustained cognitive challenge embedded in everyday classroom instruction.

Second, we show that intense cognitive challenge provides a pathway to improving academic learning. Most interventions aimed at addressing the learning crisis operate through pedagogy, curriculum reform, tutoring, mentoring, remedial instruction, or socio-emotional skills (Ashraf et al., 2021; Banerjee et al., 2007; Alan and Ertac, 2018; Alan et al., 2019; Alan and Mumcu, 2024; Duflo et al., 2024; Hassan et al., 2024; Rege et al., 2024). Related work has targeted individual cognitive skills such as working memory (Berger et al., 2025), or cognition more broadly (Brown et al., 2025). Our intervention instead directly targets neither literacy nor numeracy and crowds out some curricular instruction. Yet children become better learners while also developing stronger higher-order reasoning skills. The program is inexpensive to implement (approximately $15 per child) and can be integrated into classroom instruction, offering a scalable approach to reducing learning poverty while strengthening children's capacity to reason.

Third, we contribute to the literature on measuring higher-order skills by developing and validating a new language-based assessment of critical thinking suitable for young children. Despite its central role in decision making, problem solving, and innovation, no established instrument exists to measure critical thinking at this age. Our task captures evidence evaluation and logical inference, argument justification, and the detection of flawed reasoning. It offers a simple approach to measuring an important but rarely observed dimension of children's cognitive development.

The remainder of the paper is organized as follows. Section 2 describes the study context, the cognitive expansion program, and the evaluation design. Section 3 presents our

[3]There are several cognitive training programs aimed at improving executive functions such as attention and working memory. These studies are often small-scale and typically provide limited evidence on whether gains translate into improvements in academic learning. See Sala and Gobet (2020) for a meta-analysis of these interventions.

measurement toolkit and main outcomes. Section 4 describes the data and presents the main results. Section 5 examines potential mechanisms, first through effects on teachers and then using a dynamic model of human capital accumulation. Finally, Section 6 concludes.

2 Background

The broader study was initiated in Fall 2022 with the aim of monitoring the cognitive development of socioeconomically disadvantaged children in three southeastern provinces of Turkey: Adana, Gaziantep, and Şırnak. Within this longitudinal study, we selected two provinces, Gaziantep and Şırnak, to implement a randomized controlled trial beginning in Fall 2023. The objective of the program is to place socioeconomically disadvantaged children on a steeper development trajectory. Below, we provide background on the study context and describe the setting in which the randomized controlled trial was implemented.

2.1 Context for the Evaluation

Turkey has twelve years of compulsory schooling, with the first four years comprising primary education, during which a single classroom teacher delivers almost all subjects. It is an upper-middle-income country with a centralized, tuition-free public education system and centrally appointed teachers. While overall teacher quality is relatively strong and major infrastructural deficiencies have largely been addressed, substantial regional disparities in learning persist. According to the World Bank's 2024 report (Mundial et al., 2022), learning poverty stands at 19% nationally but rises to 35% in the disadvantaged Southeast. In response, the Ministry of Education has undertaken several reforms, most recently placing greater emphasis on the development of higher-order thinking skills and socioemotional competencies. The program we evaluate emerged against this institutional background and implemented in two southeast provinces of Turkey, Sirnak and Gaziantep.

Şırnak is a province located along the Iraqi border, where the majority of the population speaks Kurdish as a first language. Instruction in public schools, however, is in Turkish, requiring many children to acquire proficiency in Turkish while simultaneously learning to read and write during the early years of primary school. Gaziantep, located near the Syrian border, is more urban and cosmopolitan and has, in recent years, become an important industrial and economic hub in the Southeast. It is the ninth largest city in Turkey by

population. Turkish is the mother tongue for all children in Gaziantep.

Beyond differences in language environment, the two provinces also differ markedly in socioeconomic conditions and teacher labor markets. Şırnak is highly disadvantaged, with chronic teacher shortages and high teacher mobility. Many teachers accept initial appointments due to the binding nature of public service contracts, but once eligible to request a transfer (typically four years after appointment), they often seek reassignment to more desirable locations.[4] This generates substantial instability for students who are already disadvantaged in multiple domains. Additional turnover arises when school principals rely on temporary substitute teachers to fill vacancies while awaiting centrally appointed staff. In such cases, principals tend to reassign permanent teachers from higher grades (e.g., Grades 3 or 4) to lower grades (Grades 1 or 2), as early-grade instruction is considered essential for foundational skill development. Teacher turnover also occurs in Gaziantep, but at a considerably lower rate.

2.2 A Cognitive Expansion Program: Developing Brain

The program, "Developing Brain," was developed in response to persistent learning deficits and widening socioeconomic disparities in educational outcomes. The Turkish Ministry of Education has increasingly emphasized holistic child development and higher-order thinking skills as part of its national curriculum. Against this background, we collaborated with an Istanbul-based think tank and Bilkent University to develop a neuroscience-informed program for primary school students.

Developing the materials took approximately one year. The process involved repeated piloting and revision, with input from the Ministry of Education, experienced primary school teachers, neuroscientists, and artists specializing in children's educational materials. The final program combines digital and physical materials, including structured workbooks, play cards, coloring tools, and visual aids. It was also designed to work with the digital boards available in nearly all Turkish public schools.

A defining feature of the program is its high cognitive demand. The exercises are designed to challenge the cognitive limits of both students and teachers. The underlying idea is simple:

[4]Both provinces have historically low achievement outcomes, but Şırnak ranks among the lowest in the country. The province has also experienced repeated educational disruptions due to political upheaval, further exacerbating learning gaps. In addition, teacher turnover has historically been high due to perceived security risks, despite a largely stable political climate over the past decade.

cognitive capacity expands when the brain is repeatedly exposed to tasks that require effort, adaptation, and error correction. Repetition of easy or familiar tasks is not enough. The challenge must increase as capacity develops. This principle guided both the design and implementation of the program.

Teachers received extensive training at the beginning of Grade 2 and again at the beginning of Grade 3. Each training began with seminars on the neural mechanisms of learning and neuroplasticity. Teachers learned about the role of cognitive effort and challenge in strengthening neural pathways, and that neural plasticity persists beyond childhood. They then worked through the program exercises themselves. The workshops were deliberately hands-on. As the exercises became more difficult, teachers experienced the cognitive demands of the program first-hand. This helped them understand a central feature of the intervention: activities should be challenging enough to require effort, but not so difficult that students disengage.

The first year of the program was implemented in Grade 2, when children were approximately seven years old. It focused on two foundational cognitive skills: pattern recognition and categorization. Pattern recognition requires children to identify recurring structures and regularities in information and provides a foundation for abstraction and problem solving. Categorization requires them to organize objects or ideas according to shared characteristics and is an early component of scientific thinking.

The exercises started at relatively accessible levels and became progressively more difficult. Teachers were encouraged to work on both cognitive domains each week rather than completing one domain before moving to the next. The goal was to engage multiple cognitive systems simultaneously and repeatedly. This cross-training approach was intended to exploit neural plasticity and promote broader cognitive expansion.

In Grade 3, the program moved to more advanced forms of reasoning. The main targets were logical and analogical reasoning. Students worked on increasingly difficult pattern-matching exercises and learned to detect inconsistencies in visual information. These inconsistencies were initially obvious but became progressively more subtle and abstract. Analogical reasoning exercises required students to identify a relationship in one setting and apply the same relational structure in another. This ability to transfer relationships across contexts is an important component of scientific reasoning.

In addition to the workbooks and materials supplied in grade 2, new workbooks and

supplementary materials were introduced in Grade 3. In both years, teachers were encouraged to devote five to eight class hours per week to program activities. Importantly, the program did not teach curricular content. Almost all exercises were nonverbal and visual. Children worked with shapes, pictures, patterns, cards, stickers, coloring tools, and other visual materials rather than reading or numerical exercises. The program therefore did not directly train literacy, numeracy, or other academic skills.

This intensity came at a cost. Time spent on the program displaced some regular curricular instruction, particularly in Grade 3. This tradeoff was deliberate. The program rests on the hypothesis that expanding children's reasoning capacity makes learning faster and more efficient. If so, children may learn more even when they receive less direct curricular instruction.[5] We deliberately did not introduce the program in Grade 1. Children at this age are learning to read and write, and both students and teachers already face substantial cognitive and behavioral demands. We therefore chose Grade 2 as the earliest point at which children could engage intensively with the program.

Control teachers continued with business-as-usual instruction. At recruitment, they were told that participating in the study gave them a 50 percent chance of being invited to the program and its training seminars. Control teachers were promised access to all program materials once the program ended. One benefit was provided to teachers in both treatment and control schools. At the end of each school year, we gave teachers a brief report on their class, including average academic and other test scores and the class's position in the district distribution. Teachers valued this feedback, which provided an additional incentive to participate in the study.

Examples of the workbooks, exercises and photographs of classroom implementation are provided in the Online Appendix. The intervention ultimately asks a simple question. Can sustained and progressively increasing cognitive challenge expand children's capacity to reason? We hypothesize that it can. We further hypothesize that stronger reasoning capacity will make children more efficient learners, improving academic performance even though the program itself does not teach academic content and displaces some regular instruction.

[5]Many teachers initially worried that the program would leave them behind in covering the required curriculum. They were explained that the purpose of the program was to expand children's ability to learn and thereby allow them to process new material more efficiently. If successful, this could allow teachers to cover curricular material more quickly and effectively over time.

2.3 Evaluation Design

Public primary schools in Turkey are typically large and often have multiple classrooms per grade. We recruited schools with at least two classrooms in the target grade (Grade 2). Almost all contacted schools participated, although not every teacher in all willing schools volunteered. Our criterion to include a school in the RCT was to have at least two two teachers (two classrooms) volunteering to participate in the study.

The RCT covers a total of 132 schools, with 84 located in Gaziantep and 48 in Şırnak. Randomization was conducted after the collection of rich baseline data from all participating schools. Stratifying by district, our randomization routine allocated 67 school to treatment and 65 to control (43 schools to treatment and 41 to control in Gaziantep, and 24 to treatment and 24 to control in Şırnak), with the school as the unit of randomization. One control school in Gaziantep closed after baseline data collection, leaving 40 control schools in that province. In addition, one treatment school in Şırnak exited after the first year of the trial, and one treatment school closed after the second year of the trial.

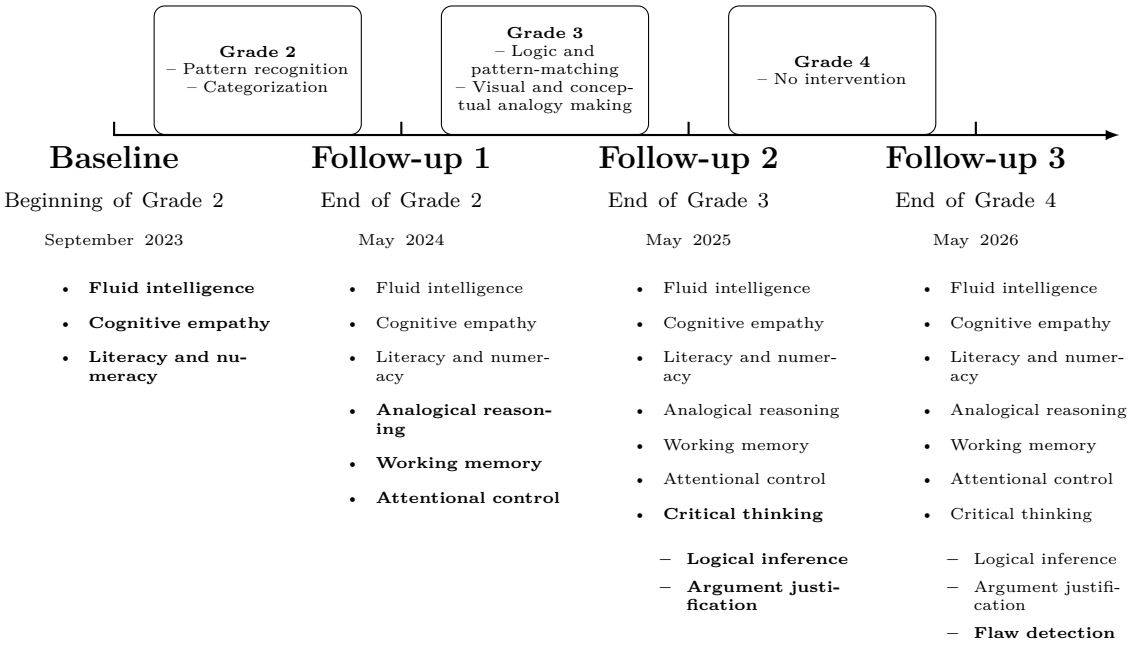
Following baseline data collection in both provinces we randomized the schools. Teachers from treatment schools were invited to the first training, implemented in September 2023 in Gaziantep and October 2023 in Şırnak. This training covered a total of 286 teachers from 67 treatment schools. At this point, the students had just started Grade 2. We collected our first follow up data in April–May 2024, at the end of Grade 2, after the first part of the program was implemented. In September 2024, treatment teachers were invited for a second training, and at this time, the students had just started Grade 3. In cases where the classroom teacher had changed, the newly assigned teacher was invited and separately briefed on the program. We made every effort to involve the new teachers, who, by definition, were not volunteers for the program. Other than one school in Sirnak, whose teachers all changed in the second year of the program, we managed to involve all new teachers.

We our second follow up data in April–May 2025, at the end of grade 3. We collected our final follow up data in April–May 2026, which marked the end of primary school for the study cohort (grade 4). No further training was provided in the final year of the study. Therefore the final follow up data allow us to estimate the persistence of the results in the absence of the program.

3 Data and Main Outcomes

In each wave, we collected data using age appropriate tests designed to capture children's cognitive development, higher order reasoning, and academic learning. At baseline, we measured fluid intelligence using Raven's Progressive Matrices (Raven and Court, 1998), a widely used measure of abstract reasoning and fluid intelligence. We also administered the Reading the Mind in the Eyes Test (RMET), which measures cognitive empathy, or theory of mind. Finally, we assessed literacy and numeracy using tests aligned with the Ministry of Education's grade specific learning objectives. Basic demographic information, including gender and date of birth, came from school administrative records.[6]

Figure 1: Timeline of program implementation and data collection



Our measurement became richer as children grew older. Figure 1 shows the progression of the assessment across the four waves. We added new measures in each wave while keeping the old ones. Fluid intelligence, cognitive empathy, attention, and working memory were measured using the same tasks across waves, allowing us to observe the developmental trajectory of skills. Academic assessments became more demanding with grade level, with increasing emphasis on reading comprehension, verbal reasoning, and mathematical reasoning. We introduced a measure of analogical reasoning in the first follow-up and made it

[6] As part of the broader longitudinal study, we also collected data on friendship networks, risk attitudes, curiosity, and growth mindset.

progressively more challenging across waves. As children became developmentally ready, we added measures of more advanced reasoning. At the end of Grade 3, immediately after the intervention ended, we introduced the first two components of our critical thinking assessment. At the end of Grade 4, we administered the complete assessment, adding the task that requires children to detect logical inconsistencies in narratives. The final wave therefore provides our most comprehensive measurement of higher order thinking.

All tests were administered in class using tablets by trained members of the research team, with support from locally recruited field assistants. We placed dividers between students to ensure privacy and independent responses. Tests were incentivized to reduce differences in effort and motivation. Below, we provide a more detailed account of our measurement toolkit.

3.1 Student Outcomes

We categorize student outcomes into four broad groups: (i)non-language-based higher-order thinking skills, (ii) core cognitive skills, (iii) academic performance, and (iv) critical thinking. Below, we describe the design of the tests (i), (ii) and (iv) and their classroom implementation.

3.1.1 Nonverbal Higher-order Skills

We measured three higher-order reasoning skills that do not rely on language competency.

- **Abstract Reasoning:** We measure abstract reasoning, which is commonly known as "fluid intelligence" using Raven's Progressive Matrices (Raven and Court, 1998). Raven's test measures the ability to identify patterns, infer relationships, and solve novel problems with minimal reliance on language or previously acquired knowledge. These abilities are closely related to the capacity to learn from new information. Given the intervention's extensive exercises in pattern recognition, visual relationships, and rotation, abstract reasoning is one of the higher-order skills we targeted directly and expected to improve. We administered Raven's test in all four waves, including at baseline. Together with baseline academic achievement, the test also provides a strong measure of children's initial ability.

- **Analogical Reasoning:** We measure analogical reasoning using a task that requires students to identify and transfer relationships across different contexts. Students are presented with multiple-choice questions based on relational mapping, where they must infer how the relationship between two items applies to a second pair. Unlike simple similarity judgments, the task requires recognizing the underlying relational structure. The test includes a range of relational forms, such as part-whole relationships, functional use, and category membership. For instance, in an item such as bee : honey :: cow : ?, students select the correct answer from a set of options (e.g., milk, hay, sheep). Choosing "milk" requires identifying the underlying producer-product relationship and transferring it to a new context. This ability to abstract relational structures and apply them across settings is a key component of higher-order reasoning. We measure analogical reasoning at the end of grades 2, 3, and 4 with progressively difficult questions. Note that the grade 2 assessment takes place before analogical reasoning is directly targeted by the program, allowing us to examine whether the first year of cognitive training generates gains in an important untargeted higher-order skill. Analogical reasoning is then directly targeted in grade 3 through progressively more demanding exercises. The grade 4 assessment, conducted one year after the intervention ended, allows us to examine the persistence of these gains.

- **Cognitive Empathy-Theory of Mind:** We measure cognitive empathy (theory of mind) using the Reading the Mind in the Eyes Test (RMET). Cognitive empathy is a socio-cognitive skill, highly correlated with other cognitive and higher order skills (Rosso and Riolfo, 2020; Baker et al., 2014). While the correlations are strong (also in our data), little is known about whether challenging higher-order thinking capacity can improve this socio-cognitive skill. RMET does not map cleanly into either language-based or non-language-based cognition. For the purposes of this study, we classify it as non-language-based because the test requires minimal reading comprehension. The task asks children to infer nuanced and sometimes abstract mental states from people's eyes. The response options consist of simple, often single-word descriptions, such as happy" or sad." Thus, the task requires little complex linguistic processing (Peterson and Miller, 2012; Rosso and Riolfo, 2020).[7] We measure cognitive empathy in all waves, including at baseline.

[7]Despite national expectations of reading proficiency, a significant number of students still struggle with reading comprehension in grade 2. Turkish is a highly phonetic language, making decoding relatively easy once children learn letter-sound correspondences. Reading comprehension can develop more slowly, however, given the complexity of its grammatical structure.

3.1.2 Executive Functioning Skills: Attention and Working Memory

We measure two core executive functioning skills: working memory and attention. Working memory allows individuals to temporarily hold information needed for an immediate task. Attention allows them to focus on relevant information while ignoring distractions. Both are widely regarded as important foundations of learning (Alloway and Alloway, 2010; Last et al., 2018; Blakey et al., 2020; Distefano et al., 2020; Zelazo and Carlson, 2020; Doebel and Lillard, 2023).

We measure working memory using a visual memory task based on (Luck and Vogel, 1997; Vogel et al., 2001; Oberauer, 2002; Halberda et al., 2006). Students see a set of colored geometric shapes on a screen for 10 seconds. The shapes then disappear, and students answer a question about their color, position, or shape. They do not know in advance which attribute they will be asked about. The task therefore requires students to hold visual information in memory for a short period.

We measure attention using the Flanker Task, a standard task in cognitive psychology (Eriksen and Eriksen, 1974). Students see a row of five arrows and must indicate the direction of the central arrow while ignoring the surrounding arrows. In congruent trials, all arrows point in the same direction (e.g., <<<<<). In incongruent trials, the surrounding arrows point in the opposite direction from the central arrow (e.g., <<><<). The incongruent trials therefore require students to ignore misleading visual information and focus on the target arrow. We implemented both two-second and one-second response-time versions of the task. Our attention measure is based on the difference in performance between incongruent and congruent trials.

Both tests were administered in all waves except baseline, as shown in Figure 1. We expect the intervention to improve both functions because the program's cognitive exercises place sustained demands on attention and working memory.[8] A full description of working memory and attention tests, including their visual structure, is provided in the Online Appendix.

[8] For example, children completed many analytical categorization exercises that required them to hold visual information temporarily in memory while completing the task. These exercises also required sustained attention. Prior research documents a strong correlation between working memory and attention, a pattern that is also evident in our data (Kane et al., 2001; Engle, 2002; Gathercole et al., 2008).

3.1.3 A Language-based higher-order thinking skill: Critical Thinking

One of the core objectives of the program was to help children develop critical thinking skills. Critical thinking requires evaluating evidence, drawing logical inferences, justifying judgments, and detecting flaws in reasoning. These skills are difficult to capture with a single measure. While survey-based instruments exist (e.g., (Alan and Mumcu, 2024)), they are not suitable for the young children in our study. We therefore developed a new task for children with basic reading comprehension.

The task has two independent parts. The first measures children's ability to evaluate evidence and draw logical inferences, as well as their ability to justify these inferences. The second measures their ability to detect flawed reasoning. We introduced the first part at the end of Grade 3. We added the second part at the end of Grade 4, as we considered it too difficult for Grade 3 students. Figure 1 shows the timing of the two parts. We describe each below.

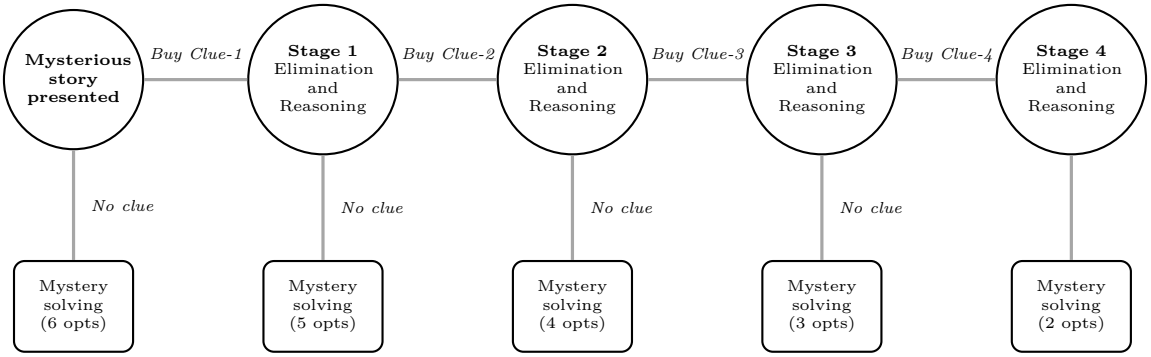
The first part of the task is an incentivized mystery-solving exercise. Students read a short paragraph about an item that has been lost in a hypothetical school. The item can be in one of six locations, but only one is correct. Students earn ten gift tokens if they identify the correct location and zero otherwise. They also receive an initial endowment of six tokens that they can use to buy clues. There are four clues, each costing one token. Each clue provides information that rules out one of the six locations. A student who purchases all four clues and uses them correctly can therefore eliminate four locations. This leaves two possible locations and a 50% chance of choosing the correct one.

The task proceeds sequentially. After reading the initial paragraph, students first decide whether to buy a clue. If they decide not to buy one, they choose among the six locations. We then ask why they chose not to purchase additional information. They select from five possible reasons: "I didn't want to spend my tokens," "I thought I could find the correct answer," "I didn't want to read more," "I thought I would not understand the clue," and "I am not curious about the correct answer." If a student buys a clue, one token is deducted from the initial endowment and the first clue appears. The clue contains information that subtly rules out one of the six locations. We then ask the student which location can be eliminated. This gives us a measure of whether the student can use the evidence to make a correct logical inference.

We next ask the student to justify the elimination. The student chooses among three

possible interpretations of the clue. The alternatives are deliberately designed to be similar, so that simply identifying the correct location is not sufficient. The student must also identify the reasoning that supports the elimination. This provides a separate measure of argument justification. The student then decides whether to buy another clue or stop and choose a location. The same process continues for up to four clues. The task therefore allows us to observe how children acquire information, use new evidence to eliminate hypotheses, and justify their reasoning. Figure 2 shows the full sequence of decisions.

Figure 2: Decision Tree for the Critical Thinking Task



Notes: The figure illustrates the sequential structure of the critical thinking task. Each student begins with an endowment of 6 tokens and a mystery story with six possible locations. At each stage, the student decides whether to purchase a clue (costing 1 token) or stop and select a location. If a student purchases a clue, they read a paragraph, identify which location is ruled out (elimination), and select the reasoning behind the elimination (justification). This sequence repeats for up to four clues. Correctly identifying the lost item's location earns 10 additional tokens. At the follow-up measurement, clue costs were removed, and all four clues were administered to every student, eliminating selection into clue-taking.

In Grade 3, our main measures of critical thinking are whether the student correctly eliminates the false option and correctly justifies this choice. These outcomes, however, are only observed for students who purchase a clue. If the treatment affects the decision to purchase clues, conditioning on this decision could introduce selection bias. We therefore designed the first round so that nearly all students were required to purchase a clue. To avoid deception, students were told that they could state their preference, but that there was a high probability that they would be required to purchase the clue and a small probability that their choice would be honored. This design allows us to observe evidence evaluation and justification for almost the entire sample in the first round. In subsequent rounds, students were free to decide whether to purchase additional clues.

We repeated the task in Grade 4 with two modifications. First, we changed the payoff structure. In the Grade 3 version, purchasing a clue was costly and therefore not payoff

maximizing. This was deliberate, as we wanted to distinguish payoff-maximizing behavior from the desire to acquire additional information, both of which could be affected by the treatment. In Grade 4, clues were free, making it payoff maximizing to take a clue. Second, we expanded the forced-clue design. In Grade 3, nearly all students were required to take a clue only in the first round. In Grade 4, students were asked whether they wanted a clue in every round, but clue taking was imposed with 90% probability regardless of their stated preference. This allows us to measure evidence evaluation and justification in all rounds, rather than only in the first round as in Grade 3.[9]

The second part of the task was implemented only in the final wave. Students read a relatively long narrative describing an ordinary everyday event. The narrative contained many details, most of which were irrelevant, and one logical inconsistency. Two statements in the narrative could not both be true. All sentences were numbered, and students were asked to identify the numbers of these two statements. The inconsistency was deliberately subtle. Students therefore had to process and retain information, distinguish relevant from irrelevant details, and detect a logical flaw in the narrative. Each student completed two such tasks.

We use the two parts of the task to construct an index of critical thinking. The index combines performance in hypothesis elimination, argument justification, and logical flaw detection. We normalize it to have mean zero and unit variance in the control group. In Grade 3, only the first part of the task was administered. The index is therefore based on hypothesis elimination and argument justification in the first round. In Grade 4, we use performance across all rounds of the first part and add logical flaw detection from the second part.[10]

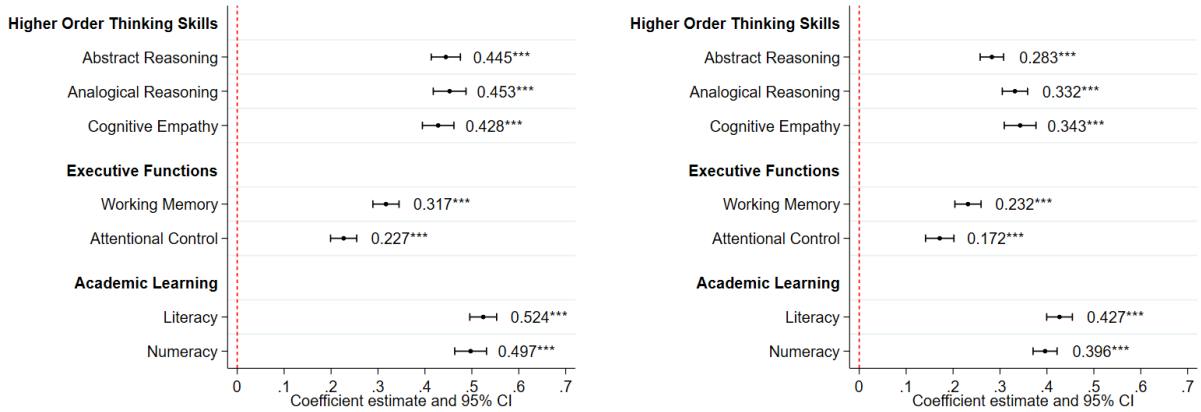
Our critical thinking index exhibits strong construct validity. Figure 3 relates the index to executive functions, nonverbal higher-order reasoning, and academic achievement in the

[9]Under this forced-clue design, a student's stated preference is honored with 10% probability. For these students, elimination and justification are not observed if they choose not to take a clue. For the remaining 90%, clue taking is imposed regardless of their stated preference, allowing us to observe elimination and justification without conditioning on the potentially endogenous decision to take a clue. Thus, we lose about 10% of the sample when identifying the causal treatment effects on these components of critical thinking. Importantly, students were informed of these probabilities in advance, so the design involves no deception.

[10]The tasks were challenging for our sample. In Grade 3, only 19% of control students correctly completed the hypothesis elimination and justification components in the first, nonselective clue round; only 5.5% correctly eliminated all four false options and justified each elimination correctly. This share was 7% in Grade 4. In Grade 4, 15% of control students correctly identified the logical inconsistency in both narratives. Only 2.5% correctly completed both components of the critical thinking assessment.

control group. Panel (a) shows the associations without conditioning on baseline fluid intelligence, while Panel (b) adds baseline fluid intelligence as a control. Critical thinking is strongly and positively associated with all three skill domains. These relationships remain sizable and statistically significant after conditioning on baseline fluid intelligence, although their magnitudes decline. This pattern suggests that performance on the critical thinking task is related to broader cognitive and academic skills, beyond being a simple reflection of children's initial fluid intelligence.

Figure 3: Predictive Validity of the Critical Thinking Index



(a) Without controlling for baseline fluid IQ (b) Controlling for baseline fluid IQ

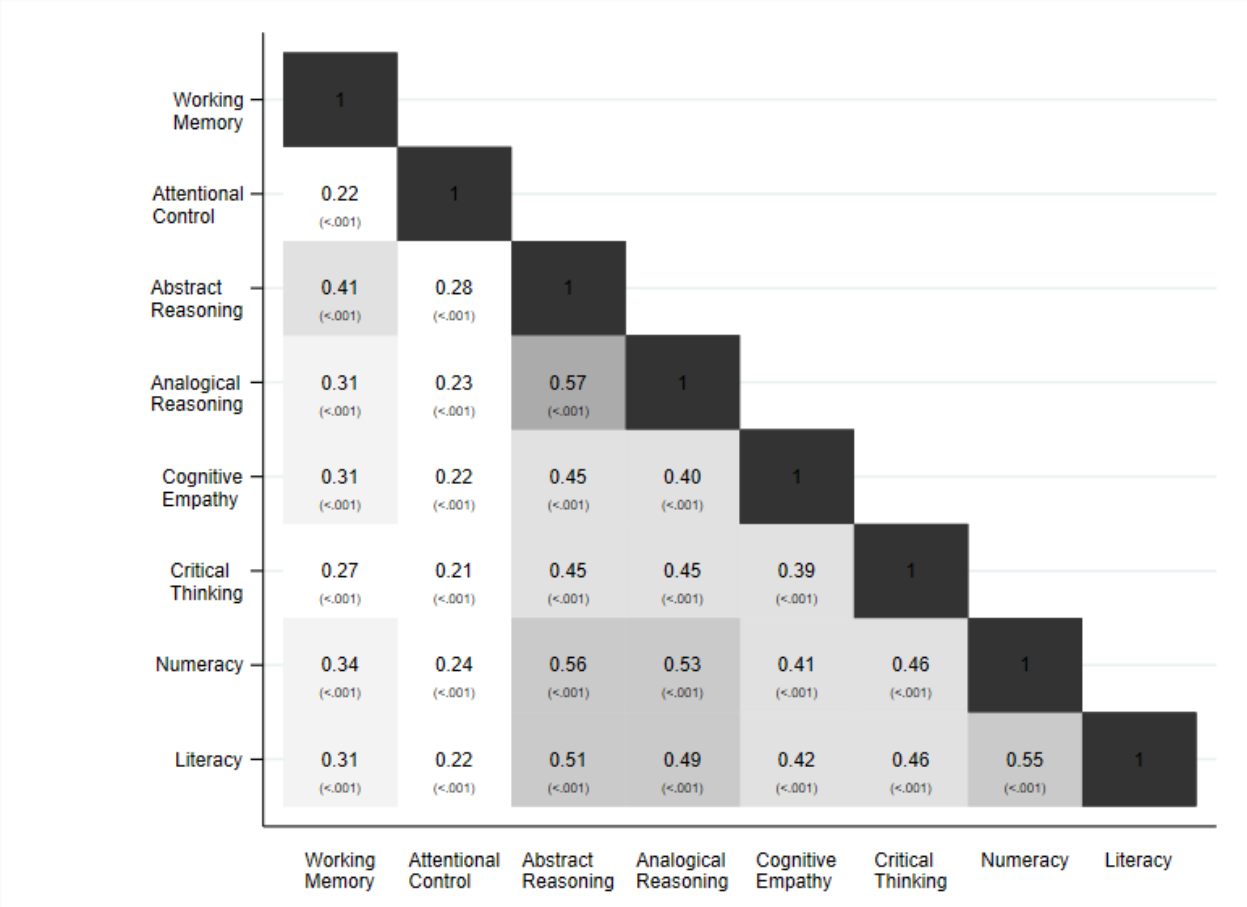
Notes: Each row reports the coefficient from a separate regression of the indicated outcome on the critical thinking index, estimated using control-group students only. All outcomes, including the critical thinking index, are standardized to have mean zero and unit variance in the control group. Panel (a) reports associations without controlling for baseline fluid IQ, while Panel (b) additionally controls for baseline fluid IQ. All regressions include district fixed effects, with standard errors clustered at the school level. Bars denote 95% confidence intervals. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

All our measures of cognitive and academic skills are strongly related. Figure 4 summarizes these associations across outcomes in the control group. The associations are positive throughout, with particularly strong relationships among the higher-order reasoning and academic measures. At the same time, the correlations are far from perfect, consistent with the measures capturing related but distinct dimensions of children's skills. These patterns provide further evidence for the construct validity of our measurement framework.

3.2 Teacher Outcomes

In addition to collecting baseline information, we administered a set of cognitive and socio-cognitive assessments to teachers. At baseline, teachers completed Raven's Progressive Ma-

Figure 4: Pairwise Associations Across Outcomes in the Control Group



Notes: Each cell reports the standardized regression coefficient from a pairwise regression of one outcome on another, absorbing district fixed effects, with standard errors clustered at the school level. P-values are shown in parentheses. The sample is restricted to the control group at follow-up (end of Grade 4). Color intensity reflects the magnitude of the association.

trices and the Reading the Mind in the Eyes Test (RMET). At follow-ups, we administered Raven, RMET, and an analogical reasoning task. When a classroom teacher changed during the study, the newly assigned teacher completed the baseline survey and the Raven and RMET assessments. We then included the new teacher in subsequent assessments.

In every wave, we also collected a rich set of teacher beliefs, attitudes, and behaviors through structured surveys. These include growth mindset (Mesler et al., 2021), motivation, industriousness, self-control, sense of responsibility, perceived competence, gender bias, critical thinking, curiosity, conformism, mental health, and perceived burnout. We also collected information on teaching styles across four dimensions: inquiry-based pedagogy, modern versus traditional teaching practices, use of extrinsic motivation, and warmth. Teachers completed all tests and surveys in a quiet room within the school. This ensured that they

were not present in their classrooms during student assessments. Example survey items for each construct are provided in the Online Appendix.

The program placed significant cognitive demands on teachers as well as students. Whether these demands can improve higher-order reasoning among adults is less clear. We therefore examine whether the program also affected teachers' abstract reasoning (Raven's Progressive Matrices), cognitive empathy (RMET), and analogical reasoning. The training may also have changed teachers' beliefs about the learning process and their teaching practices. At the same time, the program asked a lot from teachers, raising the possibility of increased stress and burnout. We examine these potential effects using our measures of teacher beliefs, teaching practices, motivation, mental health, and burnout.

4 Data and Results

4.1 Data and Internal Validity

Tables 1 and 2 present baseline balance for the student and teacher samples, respectively. For all tests and survey measures, we standardized outcomes so that the control group has a mean of zero. As shown in Table 1, the average age of students at baseline is about 88 months. Around 7 to 8 percent of students are refugees. Student characteristics are well balanced across treatment and control groups at baseline. Table 2 reports baseline characteristics for teachers. All characteristics seem well balanced across treatment and control groups. Overall, we conclude that the randomization was successful. Nevertheless, we will control for baseline cognition when estimating the impact of the program.

Table 1: Balance: Student Outcomes

	Count	Control Mean	Treatment Mean	P-value(Difference)
Student demographics				
Age in months	14994	88.21	88.11	0.30
Female (1,0)	14997	0.49	0.50	0.14
Refugee (1,0)	14997	0.08	0.07	0.55
Student cognitive skills				
Fluid IQ score	14997	0.00	0.08	0.22
Cognitive empathy score	14856	-0.00	0.05	0.35
Student academic outcomes				
Numeracy score	14739	0.00	0.08	0.26
Literacy score	14739	-0.00	0.10	0.19

Table 2: Balance: Teachers Outcomes

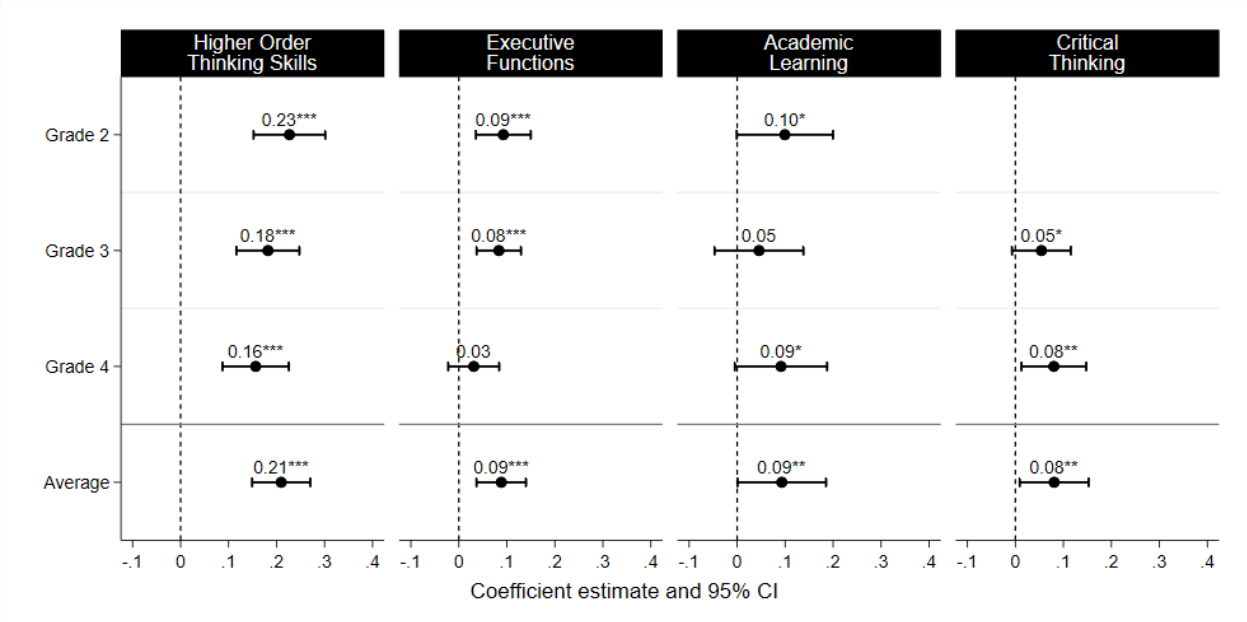
	Count	Control Mean	Treatment Mean	P-value (Difference)
Class Characteristics				
Class Size	567	34.46	34.30	0.92
Class Refugee Share	567	0.11	0.11	0.72
Teacher Demographics				
Age	567	34.03	35.01	0.25
Male (1,0)	567	0.34	0.34	0.95
Married (1,0)	565	0.56	0.61	0.27
Teacher Cognitive Skills				
Fluid IQ	566	-0.00	0.02	0.95
Cognitive Empathy	564	0.00	-0.07	0.46
Teaching Styles				
Inquiry-based Pedagogy	551	-0.00	-0.04	0.74
Modern Teaching Methods	553	-0.00	0.03	0.83
Extrinsic Motivation	551	-0.00	0.01	0.76
Emotional Warmth	554	0.00	-0.03	0.79
Teacher's Beliefs and Attitudes				
Growth-Mindset	557	-0.00	-0.03	0.84
Professional Motivation	560	-0.00	-0.13	0.22
Industriousness	555	-0.00	0.00	0.93
Self-Control	557	-0.00	0.07	0.40
Sense of Responsibility	556	0.00	0.08	0.33
Perceived Competence	562	0.00	-0.03	0.88
Gender Bias	555	0.00	-0.08	0.46
Teacher's Personality Traits				
Critical-thinking	560	0.00	0.15	0.11
Curiosity	555	0.00	-0.04	0.71
Conformism	533	-0.00	-0.06	0.57
Teacher's Personal Well-being				
Mental Health	562	-0.00	0.04	0.63
Burnout	561	0.00	-0.05	0.61

4.2 Main Results

We first present the program's effects across four broad domains: (i) nonverbal higher-order thinking skills, including abstract reasoning, analogical reasoning, and cognitive empathy; (ii) executive functions, measured by working memory and attentional control; (iii) critical thinking; and (iv) academic learning, captured by literacy and numeracy. Figure 5 traces these effects across the follow-up waves and also reports the average effect over the available post-treatment measurements.

Figure 5 shows that the program generates gains across all domains measured during the intervention. Averaged across the available follow-up waves, treatment increases nonverbal higher-order thinking skills by 0.21 standard deviations, executive functions by 0.09 standard deviations, academic learning by 0.10 standard deviations, and critical thinking by 0.08 standard deviations. The trajectories, however, differ across domains. Gains in nonverbal higher-order thinking remain sizable throughout the study period, and critical thinking, our language-based higher-order thinking measure, also remains significantly higher in Grade

Figure 5: Treatment Effects on Aggregate Skill Domains

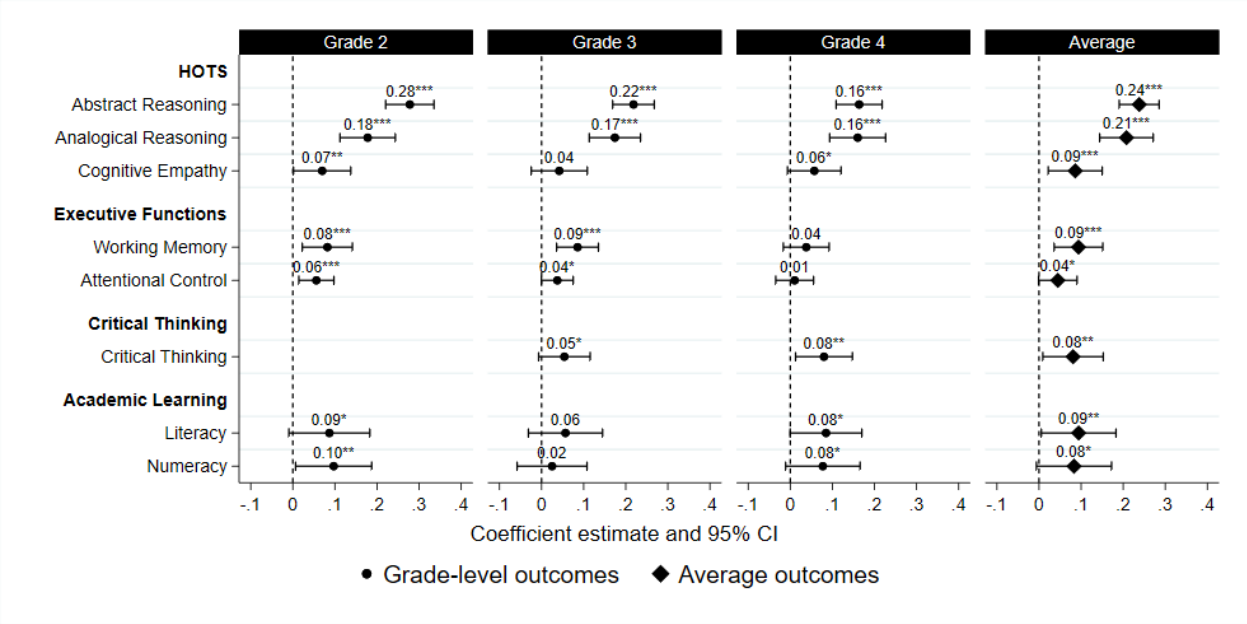


Notes: The figure reports estimated treatment effects on four aggregate outcome domains: nonverbal higher-order thinking skills, executive functions, academic learning, and critical thinking. Points denote estimated treatment effects and bars denote 95% confidence intervals. Estimates are reported separately at the end of Grades 2, 3, and 4 and, in the bottom row, averaged across the available follow-up waves. Critical thinking was first measured in Grade 3, so its average combines Grades 3 and 4. Outcomes are standardized using the control-group mean and standard deviation. Regressions control for baseline cognitive ability, a female indicator, a special-needs indicator, class size, teacher-turnover indicators for each grade, and district fixed effects. Standard errors are clustered at the school level.

4, one year after the intervention ends. This persistence across both nonverbal and verbal higher-order thinking contrasts with executive functions, where gains concentrate in the intervention years, when the program's cognitive demands are greatest, and largely disappear after the program ends. Academic learning follows yet another pattern: the effect is smaller in Grade 3, when the intervention is most likely to crowd out regular curricular instruction, but rebounds in Grade 4 and returns to roughly the initial Grade 2 gain once the program ends. Importantly, the program directly targeted only non-language-based higher-order skills through cognitively demanding exercises. The effects on executive functions, academic learning, and critical thinking therefore reflect improvements in skills not directly targeted by the program.

Figure 6 disaggregates the treatment effects into the individual skills underlying each outcome domain. Nonverbal higher-order thinking skills show sizable and persistent gains across all components. The largest effects are on abstract and analogical reasoning, the two skills most directly targeted by the program, ranging from 0.21 to 0.24 standard deviations

Figure 6: Treatment Effects on Individual Skill Measures Across Follow-up Waves



Notes: The figure reports estimated treatment effects on the individual measures underlying the four broad skill domains. Circles denote wave-specific estimates and diamonds denote averages across the available follow-up waves; bars show 95% confidence intervals. Critical thinking was first measured in Grade 3, so its average combines Grades 3 and 4. Outcomes are standardized using the control-group mean and standard deviation. Regressions control for baseline cognitive ability, a female indicator, a special-needs indicator, class size, teacher-turnover indicators for each grade, and district fixed effects. Standard errors are clustered at the school level.

on average.

The disaggregated executive-function results reinforce the pattern in the aggregate index. Working memory and attentional control improve most during the intervention years, but become smaller and statistically insignificant in Grade 4. Academic learning, in contrast, shows positive gains in both literacy and numeracy in Grade 2, weakens in Grade 3, and recovers in Grade 4. Critical thinking also improves in both waves in which it is measured. The estimated treatment effect is 0.05 standard deviations in Grade 3, when the assessment captures evidence evaluation and argument justification, and 0.08 standard deviations in Grade 4, when flaw detection is added. The more modest effect sizes in academic learning and critical thinking are consistent with the well-documented difficulty of improving verbal and language-based skills (Stanovich, 2009; Alloway and Alloway, 2010).

Overall, we find strong program effects on key cognitive outcomes. The evidence suggests that the cognitive expansion program successfully pushed children's cognitive limits, demonstrating substantial malleability in higher-order reasoning skills at young ages. Next, we

examine whether these effects vary across student characteristics and learning environments.

4.3 Treatment Effect Heterogeneity

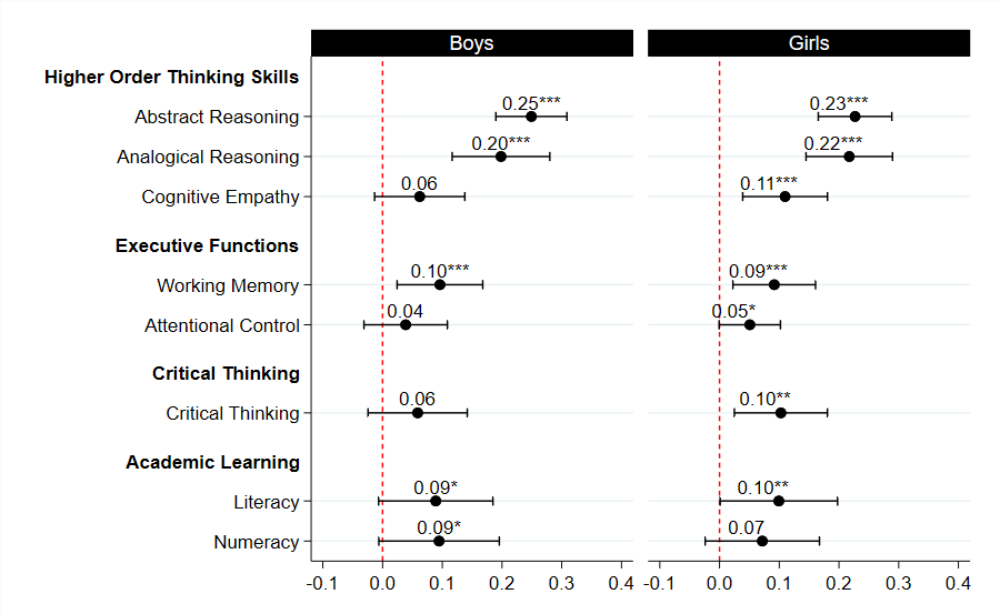
To examine whether the program had differential effects across groups, we conduct a set of heterogeneity analyses. We first study gender differences. There is substantial evidence that girls develop language skills earlier and faster than boys (Eriksson et al., 2012; Rinaldi et al., 2023).[11] Given this advantage, a cognitive expansion program may generate larger effects for girls on language-dependent higher-order skills, such as verbal reasoning, critical thinking, and socio-cognition. While all cognitive skills reinforce each other, differences in language development imply that girls' verbal development may benefit more strongly from a cognitive expansion program. As shown in Figure 7, the program generates larger effects for girls on cognitive empathy and on critical thinking, our language-based higher-order thinking measure. In contrast, treatment effects on nonverbal higher-order skills and executive functions are broadly similar across gender. Academic outcomes show a related subject-specific pattern: the estimated effect on literacy is slightly larger for girls, whereas the effect on numeracy is slightly larger for boys. Despite these differences in point estimates, none of the treatment effects differ significantly between girls and boys at conventional levels.

We then examine heterogeneity by mother tongue. Learning in one's mother tongue is known to facilitate the development of language-based higher-order skills (Cummins, 1979; Thomas and Collier, 2002). At the same time, bilingualism may also confer cognitive advantages. Prior research shows that bilingual children often exhibit stronger executive functions, including attention, task switching, cognitive flexibility, and memory, due to the sustained cognitive demands of managing two languages (Bialystok, 2015; Crivello et al., 2016). One of our study provinces, Şırnak, is predominantly Kurdish-speaking. Most children enter school with limited Turkish proficiency, and Kurdish is the dominant language at home. By comparing outcomes in Gaziantep and Şırnak, we assess whether program impacts differ between children whose mother tongue matches the language of instruction (L1) and those for whom it does not (L2).[12]

[11] Neuroscientific evidence documents earlier maturation of left-hemisphere language networks in girls, as well as faster growth in white-matter tracts associated with language processing (Wang et al., 2012; Etchell et al., 2018).

[12] Şırnak also includes a small number of non-Kurdish-speaking children, primarily children of civil servants and military personnel. Our results are robust to dropping 13 center schools that are likely to have these students.

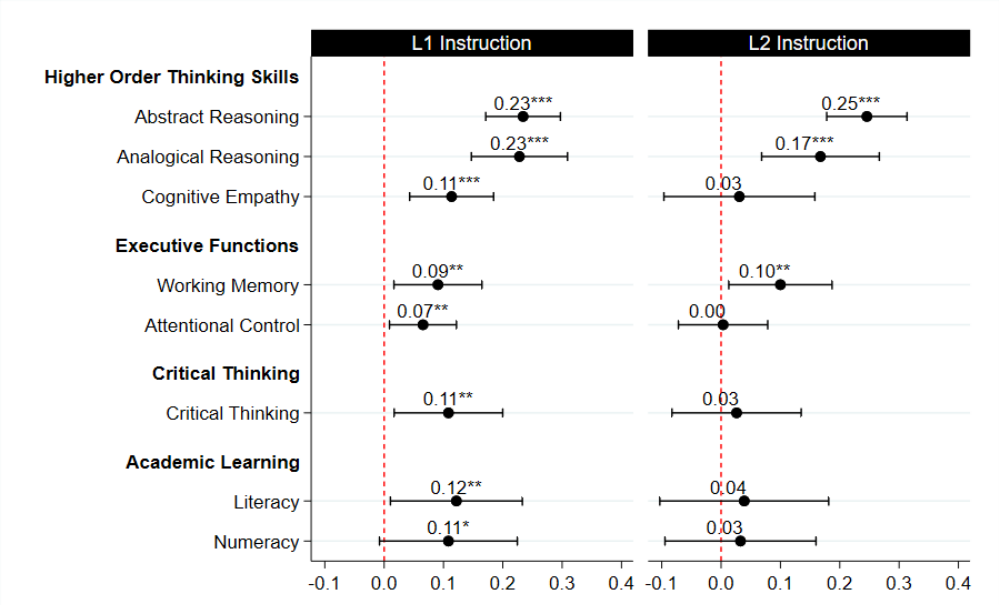
Figure 7: Heterogeneity of Treatment Effects by Student Gender



Notes: Points show estimated treatment effects separately for girls and boys; bars denote 95% confidence intervals. Outcomes are standardized using the control-group mean and standard deviation. Regressions control for baseline cognitive ability, a special-needs indicator, class size, teacher-turnover indicators for each grade, and district fixed effects. Standard errors are clustered at the school level. Stars next to coefficient estimates indicate the significance of each treatment effect; additional stars in the girls' panel indicate significant differences relative to boys.

Figure 8 presents treatment effects by language environment, comparing Gaziantep (L1) and Şırnak (L2). A striking pattern emerges. The program generates sizable and remarkably similar improvements in non-language-based higher-order skills in both environments, with substantial gains in abstract and analogical reasoning. Working memory also improves in both groups. The pattern changes, however, as outcomes become more dependent on the language of instruction. The estimated effects on cognitive empathy, critical thinking, literacy, and numeracy are considerably larger in the L1 environment and statistically significant only among students learning in their native language. In the L2 environment, by contrast, the corresponding estimates are small and statistically insignificant. Thus, the program appears to expand nonverbal reasoning capacity in both language environments, but these gains translate into academic learning and language-based higher-order thinking primarily in the L1 environment. This pattern is consistent with limited proficiency in the language of instruction constraining the extent to which improvements in nonverbal reasoning can be transferred to language-dependent skills. Despite the pronounced differences in point estimates, however, we cannot reject equality of the treatment effects across the two language environments at conventional levels.

Figure 8: Heterogeneity of Treatment Effects by Language Environment



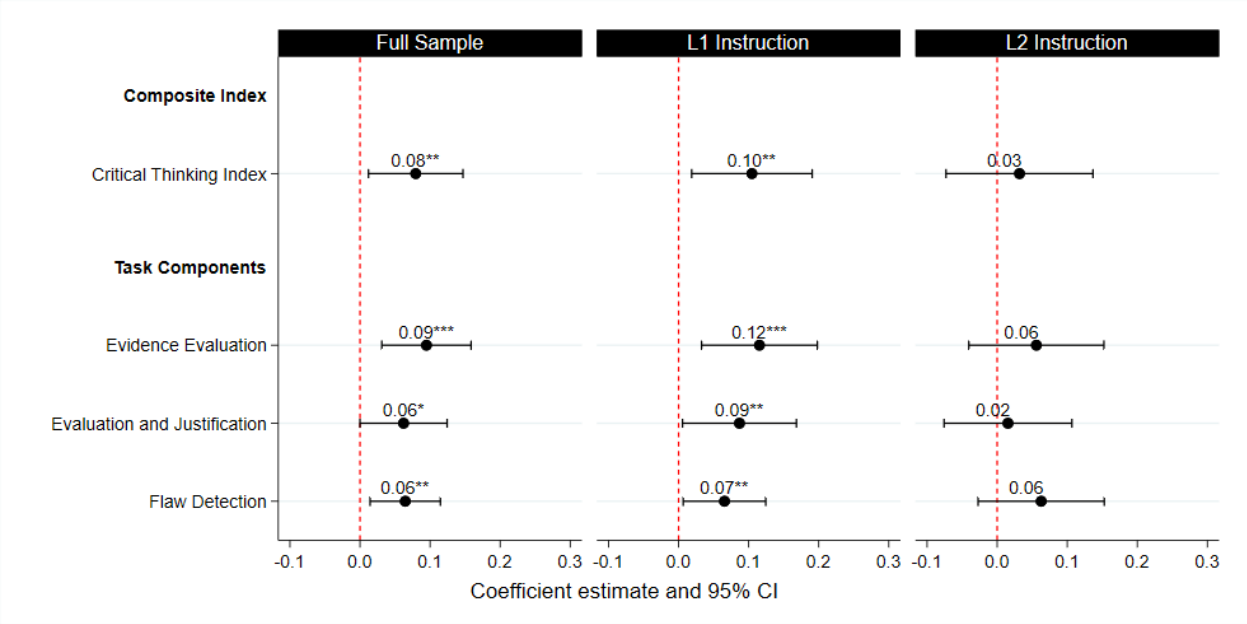
Notes: Points show estimated treatment effects separately for students in the L1 and L2 language environments; bars denote 95% confidence intervals. Outcomes are standardized using the control-group mean and standard deviation. Regressions control for baseline cognitive ability, a female indicator, a special-needs indicator, class size, teacher-turnover indicators for each grade, and district fixed effects. Standard errors are clustered at the school level. Stars next to coefficient estimates indicate the significance of each treatment effect; additional stars in the L2 panel indicate significant differences relative to the L1 environment.

Figure 9 examines this pattern more closely for critical thinking. In the L1 environment, the program improves evidence evaluation and argument justification by 0.12 and 0.09 standard deviations, respectively. The corresponding estimates in the L2 environment are smaller and statistically insignificant. The estimated effects on flaw detection, however, are similar in magnitude across the two language environments; they are significant only in L1 instruction.

Another important dimension of heterogeneity is children's baseline cognitive ability. A premise of the program is that sustained exposure to cognitively demanding activities may be particularly beneficial for children who enter school with relatively weaker cognitive skills, potentially because they have had fewer opportunities for such stimulation. Figure 10 presents treatment effects by baseline ability, measured using Raven's Progressive Matrices. Within each school, we classify students with below-median baseline Raven scores as lower ability and those at or above the median as higher ability.

The results suggest that the program was particularly effective for students who began

Figure 9: Treatment Effects on Critical Thinking Components by Language Environment

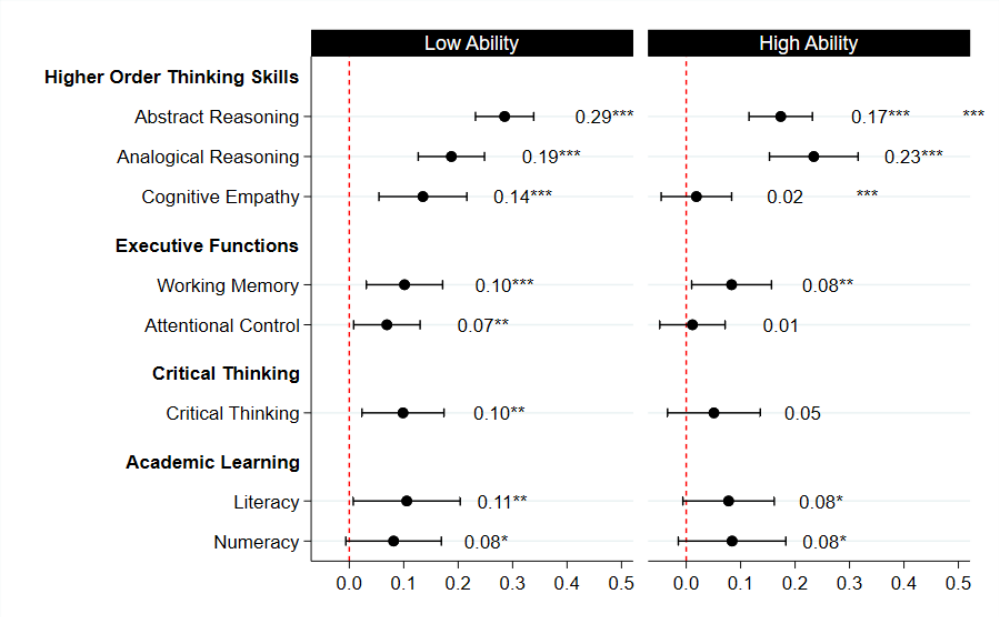


Notes: The figure reports estimated treatment effects on the Grade 4 critical thinking index and its three components for the full sample and separately for students in the L1 and L2 language environments. Points represent the estimated coefficients on the binary treatment indicator, and bars denote 95% confidence intervals. Outcomes are standardized using the control-group mean and standard deviation. Regressions control for baseline cognitive ability, a female indicator, a special-needs indicator, class size, teacher-turnover indicators for each grade, and district fixed effects. Standard errors are clustered at the school level.

with relatively lower cognitive ability within their school. This pattern is most pronounced for non-language-based higher-order skills. The effect on abstract reasoning is 0.28 standard deviations for lower-ability students, compared with 0.17 standard deviations for higher-ability students, and the improvement in cognitive empathy is also concentrated among students below the within-school median. We also find some evidence of larger effects on attention and critical thinking among these students, although differences across ability groups are not statistically significant. Overall, the results suggest that the program did not reinforce pre-existing cognitive differences within schools. If anything, the larger gains among initially lower-ability students are consistent with a narrowing of within-school differences in cognitive development.

In sum, the program generates sizable and persistent improvements in non-language-based higher-order thinking skills, as well as improvements in executive functions. The effects extend to academic learning and critical thinking, but these gains are concentrated among children who receive instruction in their native language. We also find evidence that the program was particularly effective for children who entered with relatively lower cogni-

Figure 10: Heterogeneity of Treatment Effects by Baseline Cognitive Ability



Notes: Points show estimated treatment effects separately for students with below-median and at-or-above-median baseline Raven scores; bars denote 95% confidence intervals. Outcomes are standardized using the control-group mean and standard deviation. Regressions control for baseline cognitive ability, a female indicator, a special-needs indicator, class size, teacher-turnover indicators for each grade, and district fixed effects. Standard errors are clustered at the school level. Stars next to coefficient estimates indicate the significance of each treatment effect; additional stars in the high-ability panel indicate significant differences relative to the low-ability group.

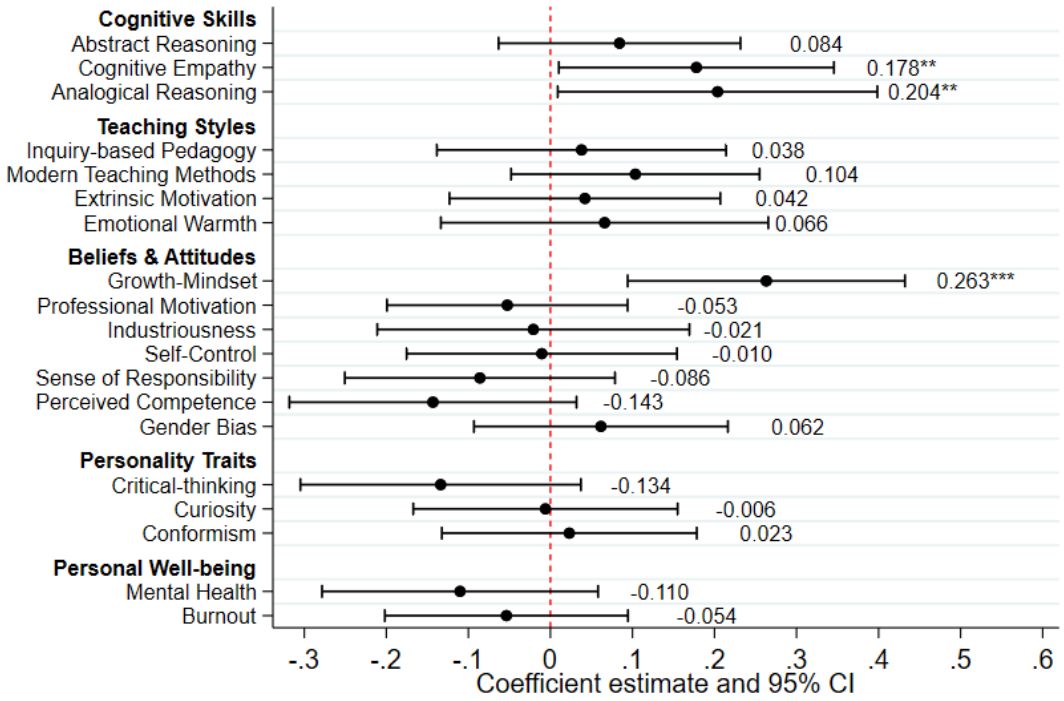
tive ability. Together, these results raise the question of how sustained cognitive challenge produces improvements across such a broad range of skills. Several potentially reinforcing mechanisms may be at work. The program may have changed teachers themselves, by strengthening their reasoning skills, changing their beliefs about how children learn, or affecting their teaching practices. At the same time, the effects may reflect dynamic complementarities in children's skill formation: stronger reasoning skills may facilitate academic learning, while learning may, in turn, support further development of reasoning skills. These channels are not mutually exclusive, as changes in teachers may strengthen the dynamic process of skill formation within children. In the next section, we first examine the effects of the program on teachers. Then, we estimate a structural model of dynamic human capital formation to better understand our results.

5 Understanding the Results: Why was the Program Effective?

5.1 Treatment Effects on Teachers

Figure 11 presents the estimated treatment effects on teachers. Although adult cognitive abilities are generally thought to be less malleable than those of children, we find that treated teachers perform significantly better on several measures. They score 0.18 standard deviations higher on the eye test and 0.20 standard deviations higher on the analogical reasoning test. The effect on Raven's test is also positive, at 0.08 standard deviations, but is not statistically significant. These improvements suggest that the sustained cognitive demands of the program affected teachers as well as students. They also point to a potential mechanism for the student effects: improvements in teachers' reasoning and socio-cognitive skills may have made them more effective in the classroom.

Figure 11: Treatment effect on teaching styles and teachers' cognitive outcomes



Notes: Each marker reports the coefficient on the treatment indicator from a separate regression of the indicated outcome on treatment, estimated on the teacher sample. Each outcome is the average of its Grade 3 and 4 measures, with each wave standardized to have mean zero and unit variance in the control group. All regressions control for the baseline value of the outcome, teacher gender, teacher experience, and endline effective class size, and include district fixed effects. Standard errors are clustered at the school level. Bars represent 95% confidence intervals. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

We find no effects on general teaching styles, but we find a substantial effect on teachers' beliefs about the learning process. Treated teachers exhibit a 0.26 standard deviation higher growth mindset, reflecting a stronger belief that intelligence and academic success can be improved through effort. Such a change may provide an additional channel through which teachers affect students, as a stronger growth mindset may influence how teachers respond to mistakes, provide feedback, and engage students (Alan et al., 2019). Finally, contrary to our initial concern, we find no evidence that the additional demands of the program increased teacher burnout. This is important given the intensity of the intervention and the substantial effort it required from teachers.

5.2 A Dynamic Model of Human Capital Formation with Higher-order Thinking Skills

Cognitive development is a dynamic process in which different skills may reinforce one another over time (Karmiloff-Smith, 1994; Welsh et al., 2010; Clements et al., 2016). The premise of our intervention is that this process can be accelerated by directly challenging children's reasoning capacity. Improvements in higher-order reasoning may facilitate academic learning, while learning may in turn further strengthen reasoning and, as language skills develop, critical thinking (Greenough et al., 1987; Knudsen et al., 2006; Diamond et al., 2007; Kolb and Gibb, 2011; Distefano et al., 2020; Slattery et al., 2022).

To study this process, we estimate a model of human capital accumulation, following Cunha and Heckman (2007, 2008); Cunha et al. (2010); Heckman et al. (2013); Attanasio et al. (2020); Agostinelli and Wiswall (2025). The model is motivated directly by the structure of the intervention. Challenging non-verbal reasoning skills may improve these skills through self-production and academic learning through cross-production. Academic learning may then further strengthen higher-order reasoning, generating dynamic complementarities between reasoning and learning. Consider the following production function for skill development:

$$\theta_{t+1,i} = f_{t+1}\Big(\theta_{t,i}, D_{t+1,i}, X_i, \eta_{t+1,i}\Big), \tag{1}$$

where $\theta_{t,i}$ and $\theta_{t+1,i}$ are vectors of child i's skills at t and $t+1$, $D_{t+1,i}$ indicates treatment during the transition from t to $t+1$, X_i is a vector of baseline characteristics, and $\eta_{t+1,i}$ is an unobserved shock to skill formation. This framework allows us to separate the direct effect

of the intervention on skill formation from the propagation of these effects across grades and skill domains.

We consider three skill domains: nonverbal higher-order thinking, academic proficiency, and critical thinking. These skills are not directly observed. Instead, we observe several assessments that provide noisy measures of the underlying skills. Using individual test scores directly would introduce measurement error and bias the estimated relationships between skills over time (Cunha and Heckman, 2008). We therefore follow Cunha and Heckman (2008) and Cunha et al. (2010) and jointly estimate the skill formation model and a measurement system that links the observed assessments to the underlying skills.

Let m^k_{jgi} denote measure j of skill $k \in \{H, A, C\}$ for child i at wave g. There are four measurement waves: baseline at the beginning of Grade 2 and end-of-year assessments in Grades 2, 3, and 4. We assume that each assessment measures one underlying skill (Gorsuch, 2003; Attanasio et al., 2020). We write:

$$m^k_{jgi} \;=\; \mu^k_{jg} \;+\; \alpha^k_{jg}\,\theta^k_{gi} \;+\; \epsilon^k_{jgi} \tag{2}$$

where θ^k_{gi} is the underlying skill, μ^k_{jg} is an intercept, α^k_{jg} is the factor loading, and ϵ^k_{jgi} is measurement error. We assume that the measurement errors have mean zero and are independent of the underlying skills. Measurement errors are independent across measures, but we allow the error of a given assessment to be correlated across adjacent waves, such that $\text{Cov}(\epsilon^k_{jgi}, \epsilon^k_{j,g+1,i}) \neq 0$. We estimate separate measurement systems for the L1 and L2 language environments. Within each language environment, we assume the same measurement system for treatment and control students.

Because the underlying skills are not directly observed, we need to normalize their scales (Anderson et al., 1956). For nonverbal higher-order thinking, we normalize the loading on abstract reasoning to one. For academic proficiency, we normalize the loading on numeracy to one. We use the same normalizations across waves. Critical thinking enters the model only at the end of Grade 4, and we normalize its scale using the hypothesis elimination component.[13] At baseline, nonverbal higher-order thinking is measured only by abstract reasoning. With only one measure, we cannot separately identify the underlying skill from measurement error. We therefore use the observed baseline abstract reasoning score, h_{1i}, as

[13]We measure nonverbal higher-order thinking using abstract and analogical reasoning, omitting cognitive empathy for simplicity.

the initial condition for higher-order thinking.

The technology of skill formation

We parameterize the production function above as a linear dynamic technology for non-verbal higher-order thinking skills and academic proficiency. Let $k \in \{H, A\}$ index the skill domain and let $-k$ denote the other skill. For transitions $g = 1, 2, 3$, we specify the technology as follows:

$$\theta^k_{g+1,i} = \delta^k_{g+1}\theta^k_{gi} + \rho^k_{g+1}\theta^{-k}_{gi} + \gamma^k_{g+1}D_{g+1,i} + X_i'\beta^k_{g+1} + \eta^k_{g+1,i}. \tag{3}$$

The parameter δ^k_{g+1} captures self-productivity, while ρ^k_{g+1} captures cross-productivity from the other skill domain. Both are allowed to vary across transitions. At baseline, θ^H_{1i} is replaced by the observed abstract reasoning score h_{1i}, as described above. The treatment term $D_{g+1,i}$ enters each transition; its interpretation across the treatment and post-treatment periods is discussed below. The vector X_i includes child sex, special-needs status, and district fixed effects.

Critical thinking is observed only at the end of Grade 4, so we do not model a lagged critical thinking stock. Instead, we relate critical thinking to the Grade-4 stocks of higher-order thinking and academic skills:

$$\theta^C_{4i} = \beta_H\theta^H_{4i} + \beta_A\theta^A_{4i} + \gamma^C D_{4,i} + X_i'\beta^C + \eta^C_{4i}. \tag{4}$$

The coefficients β_H and β_A capture the contribution of the two accumulated skills to critical thinking, while γ^C captures the treatment component not accounted for by either measured skill.

The shocks η^k_{gi} have mean zero and are independent across children. We allow the HOTS and academic shocks to be contemporaneously correlated within each wave. For identification, we constrain the HOTS innovation variance to be constant across transitions, while allowing the academic innovation variance to vary across transitions. Critical thinking has a separate residual variance.

Estimation

We estimate the measurement system in equation (2) and the skill formation technology

in equations (3) and (4) jointly by maximum likelihood.[14] We assume that the latent skills, skill innovations, and measurement errors are jointly normally distributed, subject to the covariance restrictions described above. Joint estimation accounts directly for measurement error rather than treating estimated latent skills as observed inputs in the production function (Heckman et al., 2013; Attanasio et al., 2020). In our setting, the only input to skill production other than prior skills and baseline covariates is the randomly assigned treatment.

We estimate the model separately by language-of-instruction environment: Gaziantep, where children are instructed in their first language, and Şirnak, where instruction is predominantly in a second language. This allows the skill formation technology and measurement system to differ across the two language environments. Standard errors are clustered at the school level, the unit of randomization.

Decomposing the Treatment Effects

We use the estimated technology to decompose the treatment effect on each skill into the components transmitted through the skill formation process. Let $\Delta_g^k \equiv \mathbb{E}[\theta_{gi}^k \mid D_i = 1] - \mathbb{E}[\theta_{gi}^k \mid D_i = 0]$ denote the model-implied treatment effect on skill k at wave g. Randomization implies $\Delta_1^H = \Delta_1^A = 0$. Thus, for $k \in \{H, A\}$ and $g = 1, 2, 3$,

$$\Delta_{g+1}^k = \delta_{g+1}^k \Delta_g^k + \rho_{g+1}^k \Delta_g^{-k} + \gamma_{g+1}^k. \tag{5}$$

The treatment effect on each skill therefore consists of the treatment component in that transition, γ_{g+1}^k, together with the effects propagated through own-skill and cross-skill production. For critical thinking at the end of Grade 4,

$$\Delta_4^C = \beta_H \Delta_4^H + \beta_A \Delta_4^A + \gamma^C. \tag{6}$$

At the end of Grade 2, the treatment effect on each skill is simply the corresponding treatment component, $\Delta_2^k = \gamma_2^k$. From Grade 3 onward, self- and cross-productivity propagate these effects across skill domains. For $k \in \{H, A\}$, the Grade-3 treatment effect can be written as:

[14] Estimation uses full-information maximum likelihood (FIML), so each child contributes to the likelihood through the assessments observed for that child, without requiring a balanced panel.

$$\Delta_3^k = \underbrace{\delta_3^k \gamma_2^k + \gamma_3^k}_{\text{originating in skill } k} + \underbrace{\rho_3^k \gamma_2^{-k}}_{\text{originating in skill } -k} \tag{7}$$

At the end of Grade 4,

$$\Delta_4^k = \underbrace{\left(\delta_4^k \delta_3^k + \rho_4^k \rho_3^{-k}\right) \gamma_2^k + \delta_4^k \gamma_3^k}_{\text{originating in skill } k} + \underbrace{\left(\delta_4^k \rho_3^k + \rho_4^k \delta_3^{-k}\right) \gamma_2^{-k} + \rho_4^k \gamma_3^{-k}}_{\text{originating in skill } -k} + \underbrace{\gamma_4^k}_{\text{remaining effect}} \tag{8}$$

Thus, a treatment component originating in one skill may affect either final skill through both self- and cross-productivity. The first two terms in equation (8) collect the effects originating in higher-order thinking skills and academic proficiency, respectively, while γ_4^k captures the treatment effect not transmitted through the measured Grade-3 skill stocks.

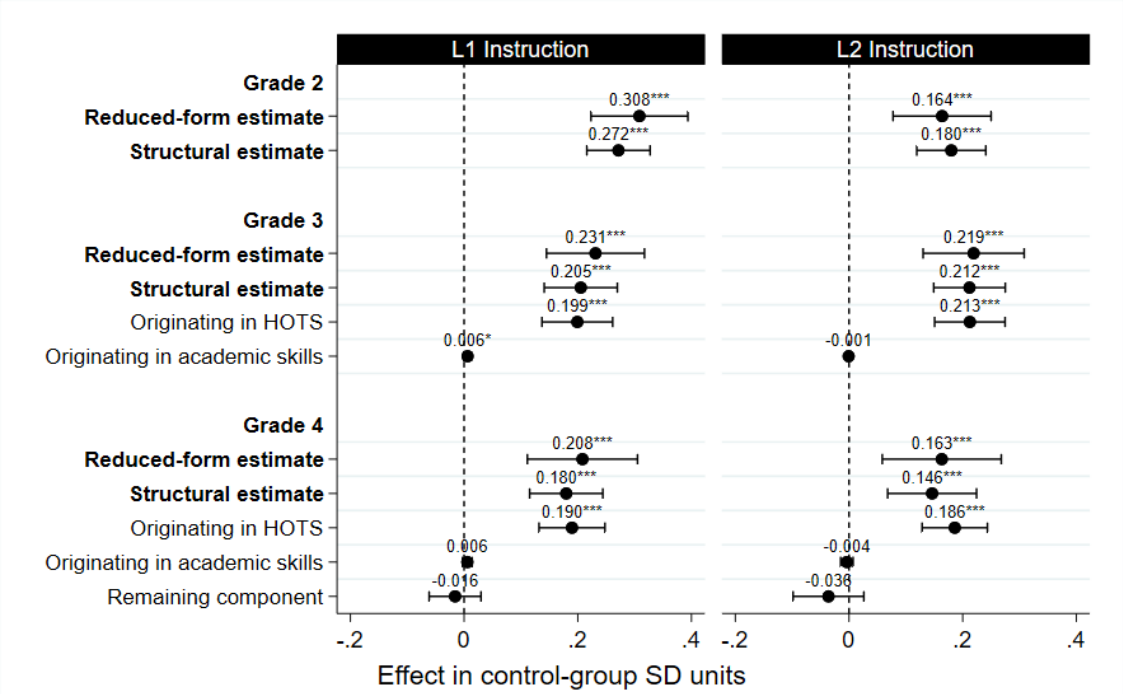
For critical thinking, we substitute the Grade-4 decompositions into equation (6). The HOTS-origin component combines the parts of Δ_4^H and Δ_4^A originating in HOTS, weighted by β_H and β_A; the academic-origin component is defined analogously. The remaining effect is γ^C.

From Grade 3 onward, each decomposition component is a nonlinear function of the estimated technology parameters. We compute point estimates and delta-method standard errors using the school-clustered covariance matrix of the maximum likelihood estimates. By construction, the components sum to the corresponding model-implied treatment effect.

Figures 12, 13 and 14 present the decomposition of the treatment effects for each skill domain. The model shows how the initial effects of the program propagate across skills and over time. The results for nonverbal higher-order thinking are very clear (Figure 12). Most of the later gains originate in the program's effects on HOTS. In the L1 environment, for example, the structural gap in Grade 3 is 0.208 standard deviations. Of this, 0.201 originates in HOTS, combining the persistence of the Grade-2 effect with the direct treatment effect on HOTS in Grade 3. By Grade 4, after the program has ended, most of the treatment effect continues to originate in the HOTS gains generated during the treatment period. The same pattern is also present in the L2 environment. Thus, improvements in higher-order thinking skills are highly persistent through self-production.

For academic skills, higher-order thinking plays an important role through cross-production (Figure 13). In Grade 3, earlier improvements in higher-order thinking skills account for the

Figure 12: Decomposition of the Treatment Effect on Higher-Order Thinking Skills

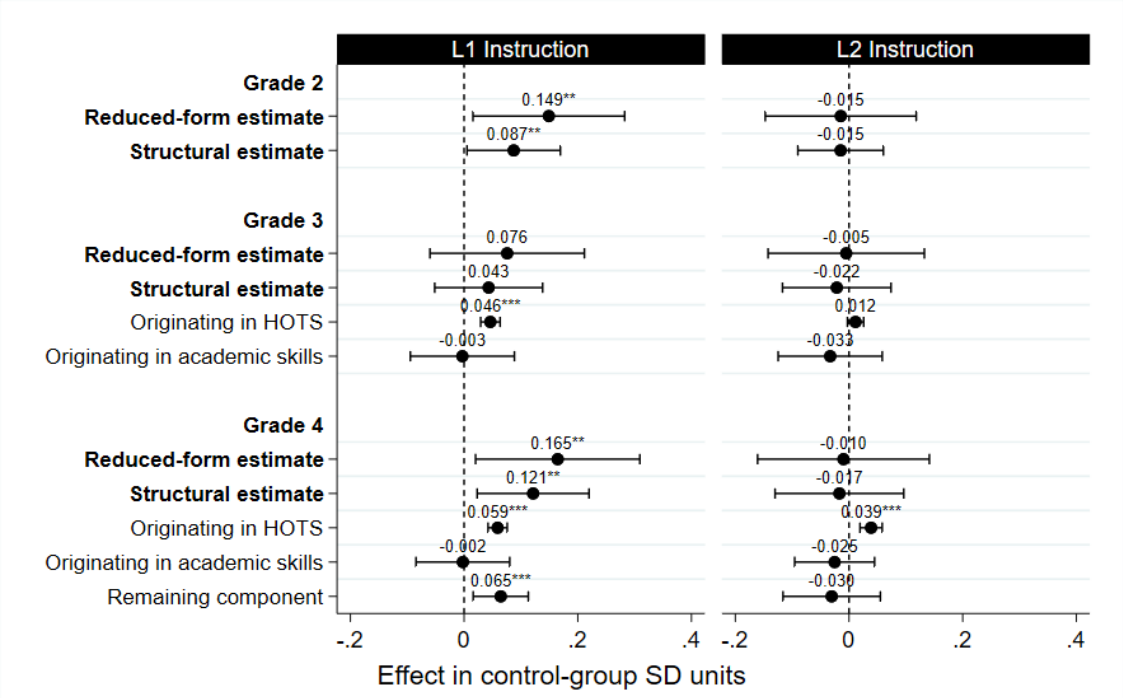


Notes: The figure compares the reduced-form treatment effects with the corresponding model-implied effects and their decomposition. The HOTS-origin component collects effects originating in the treatment components for nonverbal higher-order thinking and subsequently propagated through the skill formation technology. The academic-origin component is defined analogously for academic proficiency. In Grade 4, the remaining effect corresponds to the treatment component not transmitted through the measured Grade-3 skill stocks. Estimates are shown separately for the L1 and L2 language-of-instruction environments.

entire structural academic gain in the L1 environment. In Grade 4, effects originating in cross-production account for about half of the academic gain. The remaining half is not transmitted through the measured Grade 3 skill stocks. Interestingly, higher-order thinking also contributes positively to academic skills in the L2 environment. However, this is not enough to generate an overall improvement in academic performance.

We see a similar pattern for critical thinking (Figure 14). Here, the decomposition asks how much of the treatment gap in critical thinking at the end of primary school can be accounted for by the higher-order thinking and academic skill stocks accumulated through the program. In the L1 environment, the structural gap is 0.082 standard deviations. Half of this, 0.041 standard deviations, is accounted for by the component originating in accumulated higher-order thinking skills. The component originating in academic skills accounts for almost none of the effect. The remaining half is not accounted for by the measured skill stocks. This may reflect other effects of the program, including changes in teachers or the

Figure 13: Decomposition of the Treatment Effect on Academic Proficiency

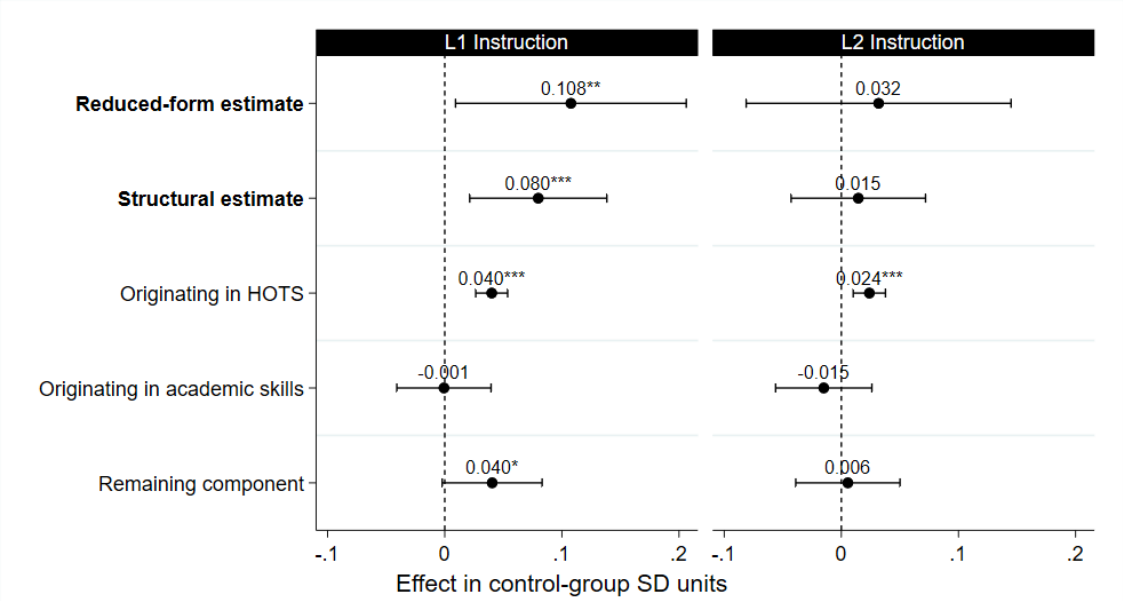


Notes: The figure compares the reduced-form treatment effects with the corresponding model-implied effects and their decomposition. The HOTS-origin component collects effects originating in the treatment components for nonverbal higher-order thinking and subsequently propagated through the skill formation technology. The academic-origin component collects effects originating in the treatment components for academic proficiency. In Grade 4, the remaining effect corresponds to the treatment component not transmitted through the measured Grade-3 skill stocks. Estimates are shown separately for the L1 and L2 language-of-instruction environments.

classroom environment.

Overall, the model highlights an important role for higher-order thinking skills. Improvements in these skills persist over time and contribute to subsequent gains in academic skills and critical thinking. At the same time, accumulated higher-order thinking and academic skill stocks do not fully account for the treatment effects, suggesting that other channels are also at work. As shown above, the program improved teachers' reasoning skills and changed their beliefs about the learning process. Improvements in children's social skills may also have altered the classroom environment in ways that facilitated learning. These channels are not captured by the model, but may help account for the remaining effects.

Figure 14: Decomposition of the Treatment Effect on Critical Thinking



Notes: The figure compares the reduced-form treatment effect with the corresponding model-implied effect and its decomposition at the end of Grade 4. The HOTS-origin component captures the contribution of treatment-induced improvements in nonverbal higher-order thinking transmitted through the Grade-4 HOTS and academic skill stocks. The academic-origin component is defined analogously for academic proficiency. The remaining effect is the treatment component not accounted for by either measured Grade-4 skill stock. Estimates are shown separately for the L1 and L2 language-of-instruction environments.

6 Conclusion

In this paper, we study whether schools can directly improve the quality of children's reasoning. We evaluate an intensive cognitive expansion program implemented over two academic years in primary schools in Turkey. The program exposes students and teachers to sustained and progressively more demanding cognitive activities. It does not directly teach literacy or numeracy and provides no remedial content. Its goal is to expand children's capacity to reason so that they become better learners.

We find significant improvements in higher-order reasoning. Treated children become better at abstract and analogical reasoning, and these gains persist one year after the intervention ends. The program also improves working memory and attention during the intervention. Importantly, these cognitive gains are accompanied by improvements in academic learning and critical thinking, despite the fact that the program crowds out some regular curricular instruction. The effects on academic learning and critical thinking, however, depend strongly on the language environment. Children improve in nonverbal reasoning regardless

of whether they learn in their first or second language, but the gains in academic learning and critical thinking are concentrated among children learning in their first language.

The effects are not limited to children. Treated teachers also improve in analogical reasoning and cognitive empathy and develop a stronger growth mindset. Our structural estimates provide further evidence on how the effects develop over time. Improvements in higher-order reasoning are highly persistent and contribute to later improvements in academic learning and critical thinking. But they do not explain all of these gains. Changes in teachers and in the broader classroom environment may provide additional channels.

Our findings also point to some important limits. Strengthening nonverbal reasoning alone may not be enough when children have limited proficiency in the language of instruction. This may be because more advanced forms of learning and critical thinking depend on verbal competency. The results therefore suggest that cognitive expansion and language development may need to proceed together.

The broader implications go beyond academic achievement. Education has traditionally placed considerable value on acquiring knowledge. But access to information is becoming increasingly cheap. This does not make education less important. It may make the ability to think about information more important. Knowing how to evaluate evidence, distinguish evidence from assertion, and detect flawed reasoning becomes particularly valuable when information is abundant and answers are easy to obtain. Schools can therefore do more than transmit knowledge. They can deliberately strengthen children's capacity to reason. This capacity matters for learning, but its value need not be measured only by test scores or future earnings. The ability to evaluate claims and reason carefully matters for decision-making and for how people engage with the world around them. Our results show that these capacities can be developed in the classroom. As knowledge becomes easier to access, learning how to think may become one of the most important goals of education.

References

Francesco Agostinelli and Matthew Wiswall. Estimating the technology of children's skill formation. *Journal of Political Economy*, 133(3):846–887, 2025.

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

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

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

Tracy Packiam Alloway and Ross G Alloway. Investigating the predictive roles of working memory and iq in academic attainment. *Journal of experimental child psychology*, 106(1): 20–29, 2010.

Theodore W Anderson, Herman Rubin, et al. Statistical inference in factor analysis. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 5, pages 111–150, 1956.

Nava Ashraf, Abhijit Banerjee, and Vesall Nourani. Learning to teach by learning to learn. *Ms., University of Chicago and Makerere University*, 2021.

Orazio Attanasio, Sarah Cattan, Emla Fitzsimons, Costas Meghir, and Marta Rubio-Codina. Estimating the production function for human capital: results from a randomized controlled trial in colombia. *American Economic Review*, 110(1):48–85, 2020.

Crystal A Baker, Eric Peterson, Steven Pulos, and Rena A Kirkland. Eyes and iq: A meta-analysis of the relationship between intelligence and "reading the mind in the eyes". *Intelligence*, 44:78–92, 2014.

Abhijit V Banerjee, Shawn Cole, Esther Duflo, and Leigh Linden. Remedying education: Evidence from two randomized experiments in india. *The quarterly journal of economics*, 122(3):1235–1264, 2007.

Eva M Berger, Ernst Fehr, Henning Hermes, Daniel Schunk, and Kirsten Winkel. The impact of working-memory training on children's cognitive and noncognitive skills. *Journal of Political Economy*, 133(2):492–521, 2025.

Ellen Bialystok. Bilingualism and the development of executive function: The role of attention. *Child development perspectives*, 9(2):117–121, 2015.

Emma Blakey, Danielle Matthews, Lucy Cragg, Jessica Buck, David Cameron, Ben Higgins, Lisa Pepper, Ellen Ridley, Emma Sullivan, and Daniel J Carroll. The role of executive functions in socioeconomic attainment gaps: Results from a randomized controlled trial. *Child development*, 91(5):1594–1614, 2020.

Christina Brown, Supreet Kaur, Geeta Kingdon, and Heather Schofield. Cognitive endurance as human capital. *The Quarterly Journal of Economics*, 140(2):943–1002, 2025.

Douglas H Clements, Julie Sarama, and Carrie Germeroth. Learning executive function and early mathematics: Directions of causal relations. *Early Childhood Research Quarterly*, 36:79–90, 2016.

Cristina Crivello, Olivia Kuzyk, Monyka Rodrigues, Margaret Friend, Pascal Zesiger, and Diane Poulin-Dubois. The effects of bilingual growth on toddlers' executive function. *Journal of Experimental Child Psychology*, 141:121–132, 2016.

James Cummins. Linguistic interdependence and the educational development of bilingual children. bilingual education paper series, vol. 3 no. 2. 1979.

Flavio Cunha and James Heckman. The technology of skill formation. *American economic review*, 97(2):31–47, 2007.

Flavio Cunha and James J. Heckman. Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources*, 43(4):738–782, 2008. doi: 10.3368/jhr.43.4.738.

Flavio Cunha, James J Heckman, and Susanne M Schennach. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931, 2010.

Adele Diamond, W Steven Barnett, Jessica Thomas, and Sarah Munro. Preschool program improves cognitive control. *Science*, 318(5855):1387–1388, 2007.

Rebecca Distefano, Erin C Schubert, Megan C Finsaas, Christopher David Desjardins, Catrina K Helseth, Marie Lister, Stephanie M Carlson, Philip David Zelazo, and Ann S Masten. Ready? set. go! a school readiness programme designed to boost executive function skills in preschoolers experiencing homelessness and high mobility. *European Journal of Developmental Psychology*, 17(6):877–894, 2020.

Sabine Doebel and Angeline S Lillard. How does play foster development? a new executive function perspective. *Developmental Review*, 67:101064, 2023.

Annie Duflo, Jessica Kiessel, and Adrienne M Lucas. Experimental evidence on four policies to increase learning at scale. *The Economic Journal*, 134(661):1985–2008, 2024.

Randall W Engle. Working memory capacity as executive attention. *Current directions in psychological science*, 11(1):19–23, 2002.

Barbara A Eriksen and Charles W Eriksen. Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics*, 16(1):143–149, 1974.

Mårten Eriksson, Peter B Marschik, Tiia Tulviste, Margareta Almgren, Miguel Pérez Pereira, Sonja Wehberg, Ljubica Marjanovič-Umek, Frederique Gayraud, Melita Kovacevic, and Carlos Gallego. Differences between girls and boys in emerging language skills: Evidence from 10 language communities. *British journal of developmental psychology*, 30(2):326–343, 2012.

Andrew Etchell, Aditi Adhikari, Lauren S Weinberg, Ai Leen Choo, Emily O Garnett, Ho Ming Chow, and Soo-Eun Chang. A systematic literature review of sex differences in childhood language and brain development. *Neuropsychologia*, 114:19–31, 2018.

Armin Falk, Fabian Kosse, Pia Pinger, Hannah Schildberg-Hörisch, and Thomas Deckers. Socioeconomic status and inequalities in children's iq and economic preferences. *Journal of Political Economy*, 129(9):2504–2545, 2021.

Sarah Flood, Joel McMurry, Aaron Sojourner, and Matthew Wiswall. Inequality in early care experienced by us children. *Journal of Economic Perspectives*, 36(2):199–222, 2022.

James R Flynn. Massive iq gains in 14 nations: What iq tests really measure. *Psychological bulletin*, 101(2):171, 1987.

Eberhard Fuchs and Gabriele Flügge. Adult neuroplasticity: more than 40 years of research. *Neural plasticity*, 2014(1):541870, 2014.

Susan E Gathercole, Tracy P Alloway, Hannah J Kirkwood, Julian G Elliott, Joni Holmes, and Kerry A Hilton. Attentional and executive function behaviours in children with poor working memory. *Learning and individual differences*, 18(2):214–223, 2008.

Richard L. Gorsuch. Factor analysis. In John A. Schinka and Wayne F. Velicer, editors, *Handbook of Psychology: Research Methods in Psychology*, volume 2, pages 143–164. Wiley, Hoboken, NJ, 2003.

William T Greenough, James E Black, and Christopher S Wallace. Experience and brain development. *Child development*, pages 539–559, 1987.

Justin Halberda, Sean F Sires, and Lisa Feigenson. Multiple spatially overlapping sets can be enumerated in parallel. *Psychological science*, 17(7):572–576, 2006.

Hashibul Hassan, Asad Islam, Abu Siddique, and Liang Choon Wang. Telementoring and homeschooling during school closures: A randomised experiment in rural bangladesh. *The Economic Journal*, 134(662):2418–2438, 2024.

James Heckman, Rodrigo Pinto, and Peter Savelyev. Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American economic review*, 103(6):2052–2086, 2013.

James J Heckman. Skill formation and the economics of investing in disadvantaged children. *Science*, 312(5782):1900–1902, 2006.

Takao K Hensch. Critical period plasticity in local cortical circuits. *Nature reviews neuroscience*, 6(11):877–888, 2005.

Michael J Kane, M Kathryn Bleckley, Andrew RA Conway, and Randall W Engle. A controlled-attention view of working-memory capacity. *Journal of experimental psychology: General*, 130(2):169, 2001.

By A Karmiloff-Smith. Beyond modularity: A developmental perspective on cognitive science. *European journal of disorders of communication*, 29(1):95–105, 1994.

Eric I Knudsen. Sensitive periods in the development of the brain and behavior. *Journal of cognitive neuroscience*, 16(8):1412–1425, 2004.

Eric I Knudsen, James J Heckman, Judy L Cameron, and Jack P Shonkoff. Economic, neurobiological, and behavioral perspectives on building america's future workforce. *Proceedings of the national Academy of Sciences*, 103(27):10155–10162, 2006.

Bryan Kolb and Robbin Gibb. Brain plasticity and behaviour in the developing brain. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 20(4):265, 2011.

Briana S Last, Gwen M Lawson, Kaitlyn Breiner, Laurence Steinberg, and Martha J Farah. Childhood socioeconomic status and executive function in childhood and beyond. *PloS one*, 13(8):e0202964, 2018.

Martin Lövdén, Lars Bäckman, Ulman Lindenberger, Sabine Schaefer, and Florian Schmiedek. A theoretical framework for the study of adult cognitive plasticity. *Psychological bulletin*, 136(4):659, 2010.

Steven J Luck and Edward K Vogel. The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657):279–281, 1997.

Rhiannon MacDonnell Mesler, Catherine M Corbin, and Brittany Harker Martin. Teacher mindset is associated with development of students' growth mindset. *Journal of applied developmental psychology*, 76:101299, 2021.

Banco Mundial et al. The state of global learning poverty: 2022 update. *World Bank Group. https://www. worldbank. org/en/topic/education/publication/state-ofglobal-learning-poverty*, 2022.

Ulric Ed Neisser. *The rising curve: Long-term gains in IQ and related measures.* American Psychological Association, 1998.

Klaus Oberauer. Access to information in working memory: exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(3):411, 2002.

Eric Peterson and Stephanie F Miller. The eyes test as a measure of individual differences: how much of the variance reflects verbal iq? *Frontiers in psychology*, 3:220, 2012.

John C Raven and John Hugh Court. *Raven's progressive matrices and vocabulary scales.* Oxford Psychologists Press, 1998.

Mari Rege, Ingunn Størksen, Ingeborg F Solli, Ariel Kalil, Megan M McClelland, Dieuwer ten Braak, Ragnhild Lenes, Svanaug Lunde, Svanhild Breive, Martin Carlsen, et al. The effects of a structured curriculum on preschool effectiveness: A field experiment. *Journal of Human Resources*, 59(2):576–603, 2024.

Pasquale Rinaldi, Patrizio Pasqualetti, Virginia Volterra, and Maria Cristina Caselli. Gender differences in early stages of language development. some evidence and possible explanations. *Journal of neuroscience research*, 101(5):643–653, 2023.

Anna Maria Rosso and Arianna Riolfo. A further look at reading the mind in the eyes-child version: Association with fluid intelligence, receptive language, and intergenerational transmission in typically developing school-aged children. *Frontiers in psychology*, 11: 586065, 2020.

Giovanni Sala and Fernand Gobet. Working memory training in typically developing children: A multilevel meta-analysis. *Psychonomic bulletin & review*, 27(3):423–434, 2020.

Eadaoin J Slattery, Eoin O'Callaghan, Patrick Ryan, Donal G Fortune, and Laura P McAvinue. Popular interventions to enhance sustained attention in children and adolescents: A critical systematic review. *Neuroscience & Biobehavioral Reviews*, 137:104633, 2022.

Keith E Stanovich. Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Journal of education*, 189(1-2):23–55, 2009.

Wayne Thomas and Virgina Collier. A national study of school effectiveness for language minority students' long-term academic achievement. 2002.

Edward K Vogel, Geoffrey F Woodman, and Steven J Luck. Storage of features, conjunctions, and objects in visual working memory. *Journal of experimental psychology: human perception and performance*, 27(1):92, 2001.

Yingying Wang, Chris Adamson, Weihong Yuan, Mekibib Altaye, Akila Rajagopal, Anna W Byars, and Scott K Holland. Sex differences in white matter development during adolescence: a dti study. *Brain research*, 1478:1–15, 2012.

Janet A Welsh, Robert L Nix, Clancy Blair, Karen L Bierman, and Keith E Nelson. The development of cognitive skills and gains in academic school readiness for children from low-income families. *Journal of educational psychology*, 102(1):43, 2010.

Philip David Zelazo and Stephanie M Carlson. The neurodevelopment of executive function skills: Implications for academic achievement gaps. *Psychology & Neuroscience*, 13(3): 273, 2020.