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

Sule Alan[†] and Merve Derebasoglu[‡]

Abstract

The quality of human thinking—our ability to abstract, evaluate evidence, and draw meaningful connections—is critical for long-run human progress. These capacities are enabled by higher-order thinking skills such as abstract, analogical, and analytical reasoning, as well as socio-cognitive reasoning. Yet the development of these higher-order skills remains underemphasized in formal education. This paper evaluates an intensive intervention designed to strengthen higher-order reasoning by deliberately pushing students' and teachers' cognitive limits over two consecutive years. The study covers 132 primary schools and follows more than 14,500 students for three years in Turkey. We assess impacts on a broad set of cognitive and higher-order reasoning skills. Central to our measurement system is a set of language-based tasks designed to capture three key components of critical thinking: the ability to evaluate evidence and draw inferences, the ability to justify one's reasoning, and the ability detect flawed reasoning. We find that the program significantly improves memory and attention, as well as non-language-based higher-order reasoning skills such as abstract, analytical, analogical, and logical reasoning for both boys and girls. In contrast, gains in language-based higher-order reasoning are concentrated among girls. Treated girls exhibit stronger critical thinking processes than untreated girls. Our results suggest that sustained exposure to cognitively demanding instructional activities can significantly improve the quality of students' reasoning and serve as a pathway to higher academic performance.

JEL Codes:

Keywords: higher-order reasoning, critical thinking, 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 Bilkent University Email: salancrossley@gmail.com.

[‡]European University Institute.

1 Introduction

"*Much of our thinking, left to itself, is biased, distorted, partial, uninformed or down-right prejudiced. Yet the quality of our life and that of which we produce, make, or build depends precisely on the quality of our thought.*" Paul and Elder (2007).

Quality of human thinking is unparalleled relative to other species. Humans can abstract, draw analogies, categorize information, and critically evaluate evidence. These capacities enable complex problem solving, creativity, and innovation. Beyond basic cognition, humans possess higher-order reasoning skills that allow them to apply logic, use language flexibly, and interpret social situations in order to make effective decisions. These skills develop naturally as the brain takes advantage of neural plasticity, which is strongest in childhood but persists well beyond it (Knudsen, 2004; Hensch, 2005; Heckman, 2006; Lövdén et al., 2010; Fuchs and Flügge, 2014). However, formal education can accelerate their development by exposing children to structured learning environments that regularly require reasoning and problem solving (Flynn, 1987; Neisser, 1998).[1] This role may be particularly important for socioeconomically disadvantaged children, who tend to receive less cognitive stimulation at home. Despite this potential, formal schooling often prioritize test scores as a proxy for the quality of human capital, rather than explicitly cultivating the cognitive processes that underlie advanced reasoning (Jackson, 2018).

This paper studies whether the quality of children's reasoning can be improved in the formal school setting. We examine the development of higher-order thinking skills among socioeconomically disadvantaged children through an intensive and lengthy in-school cognitive expansion program. The program deliberately pushes the cognitive limits of both students and teachers over two consecutive years, taking advantage of the high neural plasticity of the developing brain in the early primary school years. It is a cognitive challenge program that aims to strengthen both core cognitive (executive) skills—considered pillars of learning—and higher-order reasoning skills such as abstraction, logical reasoning, and analogical reasoning, as well as sociocognition and critical thinking. The program, called Developing Brain, consists of five modules, each contained in a separate activity book, prepared based on neuro-scientific insights by experts. Three modules are implemented in classrooms by students' own teachers in grade 2 and two in grade 3. The first modules focus on pattern recognition, categorization of objects, and working memory. The later modules focus on

[1] A note on Flynn effect? How schooling was pivotal.

analogical reasoning and pattern matching.

The study took place in two provinces in Turkey and covered 132 primary schools and more than 14,500 students and over 550 teachers. We use a clustered randomized design to roll out the program where the unit of randomization was school. Teachers in treated schools received intensive training at the beginning of grade 2 and again at the beginning of grade 3 to learn the materials themselves. The training and the subsequent implementation was demanding for teachers. After extensive workshops where teachers mastered the materials combined into 5 main activity books and supplementary materials, they were asked to deliver them to students for at least one hour each day over two consecutive academic years. Teachers were given occasional support when the material got difficult. They were instructed that the key to the success of the program is to be able to challenge the children with the right dosage. Too easy would not expand the capacities, too hard would discourage them. To convey this important idea, the teachers received a lengthy training on the neural mechanisms of learning and brain plasticity from experts.

The program did not crowd out much of the core teaching in grade 2, as the grade 2 curriculum is relatively light. However, some crowding out occurred in grade 3. This was difficult to avoid, as the program required teachers to deliver cognitively demanding material to both students and themselves, but also partly deliberate. Our conjecture was that although teachers might initially struggle to cover both the program material and the official curriculum, the cognitive expansion, if it is achieved, would ultimately accelerate learning of the official curriculum. As a result, treated students would accumulate stronger foundational literacy and numeracy skills over time along with better reasoning skills.[2]

We evaluate this intensive program using a broad set of outcomes over 3 years. We measured core cognitive skills, which are working memory and attention, and higher-order reasoning skills, both language-based and non-language-based. We collected four waves of data to assess the effectiveness of the program. Baseline data were collected at the beginning of grade 2 (September 2023). The first assessment testing (midline) took place at the end

[2]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 clearly 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 core cognitive capacities and higher-order reasoning could accelerate foundational learning. The program appears to have succeeded in this premise and is now being incorporated into the reformed Turkish national curriculum, which places less emphasis on rote factual learning and greater emphasis on cognitively demanding activities.

of grade 2 to evaluate the first phase of the program. The second assessment (endline) was conducted at the end of grade 3 to measure the full impact after both years of the intervention was completed. Finally, we collected data at the end of grade 4, which marks the end of primary school in Turkey, to assess the persistence of the effects one year after the program ended (follow-up). We increased the number of tests and their complexity over time as children grew older. We measured fluid IQ, socio-cognitive skills (cognitive empathy and norm-based reasoning), and verbal and numerical ability at baseline and at every subsequent wave. At the end of grade 2, we additionally measured attention, working memory, and analogical reasoning.

At the end of grade 3, we expanded the measurement inventory further. In addition to the existing measures, we introduced measures of risk attitudes and critical thinking processes. The latter represents one of the methodological contributions of the paper. Our aim was to assess whether sustained cognitive demands placed on children could stimulate their critical thinking processes, even without explicitly targeting them. To do so, we designed a novel language-based task that captures two key components of critical thinking: (i) the ability to evaluate evidence and draw inferences, and (ii) the ability to justify one's reasoning after making a judgment. We implemented this task as an incentivized mystery-solving game at the end of grade 3, when students were expected to have sufficient reading comprehension skills. In grade 4, we added a third component of critical thinking: the ability to detect subtle flaws or inconsistencies in arguments. This task involved reading short stories, each containing a subtle logical inconsistency. We combined performance across these tasks into a single index that captures the core elements of critical thinking and allows us to assess whether the cognitive expansion program improves these processes. We validated our measure of critical thinking using the rich set of cognitive measures collected in the study. Our critical thinking index, which combines the three components of the critical thinking process, is positively and significantly correlated with working memory and attention, as well as with academic performance and other higher-order reasoning measures, including socio-cognitive skills such as cognitive empathy.

We measured academic performance in every wave, including baseline, focusing on foundational literacy and numeracy. We designed and piloted our tests to align with the grade-level curriculum in Turkey. In grade 2, we used general foundational assessments inspired by Ministry of Education (MoE) test banks. In grades 3 and 4, we introduced more demanding academic tests that capture complex verbal and mathematical reasoning, complemented by questions on transitivity and Bayesian reasoning. These advanced verbal and numerical

tasks also reflect the expected learning gains in the recently reformed Turkish curriculum for grades 3 and 4. As such, they capture not only academic achievement but also aspects of advanced learning that are closely linked to children's higher-order reasoning skills.

We find that the program significantly improved two executive functions, working memory and attention.[3] It also substantially enhances non-language-based higher-order reasoning skills, such as abstract reasoning (0.26 sd), analogical reasoning (0.22 sd), analytical reasoning (0.18 sd) and logical reasoning, (0.06 sd) for both boys and girls, and for students whose mother tongue is either Turkish (official instruction language) or Kurdish (language of the largest minority in Turkey). In contrast, effects on language-based higher-order reasoning follow a clear pattern by gender and language environment. Students whose mother tongue is not the official language of instruction (an L2 environment) do not show significant improvements in critical thinking or verbal reasoning relative to the control group. However, our design is powered enough to show that girls in an L1 environment show clear gains in these skills. Treated girls score 0.09 standard deviations higher in critical thinking than untreated girls. They also score 0.11 and 0.09 standard deviations higher in cognitive empathy and reading comprehension, respectively.

Implementation placed significant cognitive demands on teachers. They had to master the material themselves, which progressively became more difficult even for adults. While we did not expect large improvements in adult cognition, we find evidence that the program improved several teacher skills, which may explain part of the student-level effects. Treated teachers score significantly higher in analogical and abstract reasoning than teachers in control schools. This effect is driven mainly by younger teachers, suggesting that the window for cognitive expansion may extend into early adulthood. Another pathway from teacher improvement to student outcomes may be the adoption of a growth mindset. Treated teachers are more likely to believe in neural plasticity and in the possibility that intelligence can be expanded through effort. This mindset may lead teachers to invest more in students' cognitive potential and to persist in supporting them when learning becomes challenging.

To explore mechanisms further, we estimate a dynamic human capital accumulation model, 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, which targets multiple cognitive capacities simultaneously

[3]These two cognitive functions are known as executive functioning skills that are important pillars of learning.

rather than sequentially. In particular, the program directly challenges executive functioning, pattern recognition, categorization, and analogical reasoning. We therefore model the intervention as shifting several dimensions of cognition at once. These initial gains improve children's ability to learn the official curriculum, reflected in stronger verbal and numerical ability. Academic learning and continued cognitive challenge then further expand higher-order reasoning. This creates a dynamic process in which cognition, learning, and reasoning reinforce one another over time. We estimate these relationships in a dynamic framework and show that the simulated skill trajectories generated by the model closely match the patterns in the data.

Our paper makes three contributions to the education and human capital literature. At a broad level, we take a different approach to improving learning. Rather than directly targeting test scores or specific academic content, we target the quality of children's thinking processes. We view improvements in reasoning, abstraction, and critical evaluation as a pathway to more effective learning and, ultimately, to higher-quality human capital. First, we provide causal evidence that the quality of children's thinking can be improved through the school environment. Using a large-scale cognitive expansion trial, we show that sustained exposure to cognitively demanding instruction strengthens both core cognitive capacities and higher-order reasoning among socioeconomically disadvantaged children. 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 directly targets these gaps by exposing children to the processes that underlie complex thinking—such as abstraction, analogy, logical reasoning, and critical evaluation. While cognitive science documents correlations between these skills and academic success, evidence on whether they can be deliberately developed at scale in school settings remains limited.[4] Our results show that these skills are malleable and can be improved through structured classroom activities when interventions leverage the high neural plasticity of early childhood.

Second, we provide evidence that improving the quality of thinking can serve as an effective pathway to learning. Most existing interventions aimed at addressing the learning crisis focus on pedagogy, curriculum, mentoring, remedial instruction, or socio-emotional skills, and typically target test scores directly (Banerjee et al., 2007; Alan and Ertac, 2018;

[4]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.

Alan et al., 2019; Alan and Mumcu, 2024; Duflo et al., 2024; Hassan et al., 2024; Rege et al., 2024) (Lots of ECON references, including my curiosity, grit, some mentoring papers. In contrast, we target the underlying cognitive processes that support learning. Other examples, Ashraf et al. (2021) pedagogical intervention aiming at improving critical thinking, Berger et al. (2025) paper targeting working memory, Brown et al. (2025) cognition. We show that deliberately pushing students' cognitive limits improves not only core cognitive skills but also higher-order reasoning abilities that are central to critical thinking and innovation. These gains translate into stronger academic performance, suggesting that improving how children think can enhance how they learn. Our approach complements existing strategies and offers an alternative pathway to improving foundational learning.

Third, we contribute methodologically by developing and validating a new measure of critical thinking suitable for young children. Our language-based tasks capture key components of the critical thinking process, including evaluating evidence, drawing inferences, justifying reasoning, and detecting flawed arguments. We combine performance on these tasks into a single index and show that it is strongly correlated with core cognitive outcomes, academic skills, and other higher-order and socio-cognitive abilities. This provides a scalable way to measure the quality of thinking in young populations. Together, these contributions suggest that targeting the quality of thinking is a promising and cost-effective approach to improving learning and reducing early human capital gaps. The program is inexpensive to implement (costing $15 per child) and, if integrated into existing curricula, can provide a scalable way to foster stronger, more capable, and more adaptable learners.

The remainder of the paper is organized as follows. Section 2 summarizes the context in which the program was implemented and its key features. Section 3 details the evaluation design. Section 4 outlines our extensive toolkit we use to evaluate the program. Section 5 provides details of our the data and presents our main results. Section 6 discussed the mechanisms via a dynamic human capital accumulation model. We conclude in Section 7.

2 Background

The broader study was initiated in Fall 2022 with the aim of monitoring the cognitive, socio-cognitive, and socio-emotional development of socioeconomically disadvantaged children in three southeastern provinces of Turkey: Adana, Gaziantep, and Şırnak. Within this longitudinal framework, we selected two provinces—Gaziantep and Şırnak—to implement

a randomized controlled trial beginning in Fall 2023. The trial evaluates the effectiveness of a program designed to strengthen children's higher-order reasoning skills by deliberately pushing their cognitive limits through neuroscience-informed instructional exercises. The overarching objective of the program is to place socioeconomically disadvantaged children on a steeper cognitive and socio-emotional 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 quality 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. Our program 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 conducted 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 except for Syrian refugees 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.[5] 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), reflecting the critical importance of early-grade instruction 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 growing concerns about persistent learning deficits and widening socioeconomic disparities in educational outcomes. In recent years, the Turkish Ministry of Education has placed increasing emphasis on holistic child development, including a revision of national curricular priorities toward higher-order thinking skills, which are now widely viewed as central to the quality of human capital and, ultimately, to a country's capacity for innovation and long-run competitiveness. In line with these priorities, we collaborated with an Istanbul-based think tank and Bilkent University to develop a neuroscience-informed educational program for primary school students.

The material development process lasted approximately one year and involved multiple rounds of piloting and revision, incorporating feedback from the Ministry of Education, experienced primary school teachers, neuroscientists, and artists specializing in children's educational materials. The final program was designed to be fully compatible with the digital boards available in nearly all Turkish public schools and included a rich set of digital and physical materials, such as structured workbooks, play cards, coloring tools, and visual aids.

A defining feature of the program is its deliberately high cognitive demand for both students and teachers. This design choice reflects a central premise from cognitive neuroscience: expanding cognitive capacity through neuroplasticity requires sustained exposure to challenging tasks that push the limits of existing cognitive capacities. Cognitive growth does

[5]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.

not occur through repetition of familiar and easy material but through progressive challenge that requires effortful engagement and error correction. This principle was made explicit throughout the design and emphasized during the implementation of the program.

Teachers received extensive training at the beginning of Grade 2 and again at the beginning of Grade 3. These trainings were structured as intensive, hands-on workshops in which teachers worked through the exercises themselves before implementing them in the classroom. Prior to working on the materials, teachers attended seminars on the neural mechanisms of learning and neuroplasticity. They learned the role of cognitive effort and challenge in strengthening neural pathways at any age. During the workshops, teachers experienced first-hand the demanding nature of the exercises and developed an understanding that the program required sustained cognitive effort from both themselves and their students.

The first year of the program, implemented in Grade 2 (approximately age 7), targeted three foundational cognitive skills: working memory, pattern recognition, and categorization (classification). Working memory, the ability to temporarily hold and manipulate information necessary for complex tasks, is a core component of executive functioning and a key determinant of learning capacity. Pattern recognition involves identifying recurring structures and regularities in information and underlies problem solving and abstraction. Categorization refers to organizing objects or ideas into groups based on shared characteristics and constitutes an early foundation of scientific thinking. During Grade 2, students engaged in a wide range of exercises targeting these three skills. Importantly, the exercises were sequenced to begin at relatively accessible levels and became progressively more demanding over time. Teachers were encouraged to cover all three skill domains within each week rather than focusing on a single domain in isolation. This cross-training approach was motivated by evidence that simultaneous engagement of multiple cognitive systems best exploits neural plasticity and promotes broader cognitive expansion.

In the second year of the program, implemented in Grade 3, the focus shifted toward higher-order reasoning skills, specifically logical reasoning and analogical reasoning. Logical reasoning was trained through increasingly demanding pattern-matching exercises and tasks requiring students to detect inconsistencies—initially obvious ones and, over time, more subtle and abstract contradictions. Analogical reasoning tasks required students to identify relational structures in one context and transfer them to another, a precursor for scientific thinking. New workbooks and supplementary materials were introduced, and teachers were encouraged to devote between three and four hours per week to program activities in both

the first and the second year of the program. It is important to emphasize that none of the exercises were language-based, i.e., none involved any reading materials. The materials are shapes, pictures, stickers and coloring tools.

The intensity of the program inevitably crowded out some time that would otherwise have been allocated to core curricular instruction. This was anticipated by design. The program rests on the premise that by expanding children's cognitive capacity, learning itself becomes faster and more efficient. Under this view, temporary reductions in instruction time are offset by gains in the quality and depth of learning.[6] For this reason, the program was intentionally not introduced in Grade 1, when children are learning to read and write and cognitive and behavioral demands are already extremely high on them and their teachers. Grade 2 was chosen as the earliest point at which children could engage meaningfully with both language-based and non-language-based higher-order cognitive tasks.

Examples of exercises for each targeted skill, along with snapshots of classroom implementation, are provided in the Online Appendix. By placing sustained and progressively increasing cognitive demands on students and teachers, the program is designed to expand children's cognitive capacity, strengthen higher-order reasoning, including abstract, analytical, logical, and analogical reasoning, and, over time, foster more advanced verbal reasoning and critical thinking processes. We hypothesize that these changes will not only enhance cognitive functions directly but also translate into more efficient learning and improved academic outcomes, despite the temporary crowding out of standard instruction.

3 Evaluation Design

Public primary schools in Turkey are typically large and often have multiple classrooms per grade. We therefore recruited schools with at least two classrooms in the target grade (Grade 2). Participation in the study required teachers' willingness to be part of the study and it was conveyed to the teachers that the program was demanding but their participation was voluntary. 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

[6]Many teachers initially raised concerns about the possibility that the program would crowd out core curricular instruction and leave them behind in terms of required content coverage. These concerns were answered by emphasizing that expanding cognitive capacity—particularly working memory and attentional endurance—could ultimately allow students to process information more efficiently, enabling teachers to cover curricular material more quickly and effectively over time.

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 the 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 and randomization, 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. The midline data collection was conducted 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 again 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. The second endline data collection was carried out in April–May 2025, at the end of grade 3. No further training was provided in the final year of the program. The final follow up was conducted in April–May 2026, which marked the end of primary school for the study cohort. The final wave (follow-up) provides us with the persistence of the results in the absence of the program. Figure OA.0.2 in the Online Appendix presents the timeline of the randomized controlled trial.

4 Data and Main Outcomes

In each wave, we collected data using age-appropriate tests designed to capture children's cognitive and higher-order reasoning development. At baseline, we measured fluid intelligence using Raven's Progressive Matrices (Raven and Court, 1998), a widely used test in cognitive science and economics that captures abstract reasoning and learning ability. We

also administered the Reading the Mind in the Eyes Test (RMET), which measures cognitive empathy (or theory of mind). In addition, we implemented a task measuring analytical thinking, referred to as the Triad task, which we describe in detail below.

These cognitive and socio-cognitive measures were repeated in all subsequent waves. In addition, we administered literacy and numeracy assessments aligned with the Ministry of Education's grade-specific learning objectives. After baseline, the academic assessments were designed to emphasize reading comprehension in verbal tests and logical reasoning in mathematics tests. Finally, in each wave, including baseline, we collected social network data by asking children to nominate their best friends within the classroom, applying a soft cap of five nominations per student. Basic demographic information, including gender and date of birth, was obtained from school administrative records.

As children progressed to higher grades and became developmentally ready, we expanded our assessment toolkit by introducing additional measures, such as analogical reasoning and critical thinking. Online Appendix Table OA.0.1 summarizes the full set of outcome measures administered in each wave. Below, we provide a detailed description of the main outcome variables measured at midline, endline, and follow-up. These outcomes capture the cognitive and higher-order thinking skills for which we expect the intervention to have impacts on both students and teachers.

4.1 Student Outcomes

We categorize outcomes into three broad categories: (i) core cognitive skills, (ii) non-language-based higher order thinking skills, and (iii) language-based higher order thinking skills. All tests were implemented in-class, by highly trained academic team members using tablets, with the help of field assistants. Before the tests, we placed divisions between students to ensure privacy. Below, we explain in detail how we design these tests and how we implemented in the classrooms.

4.1.1 Core Cognitive Skills: Attention and Working Memory

We measure two core executive functioning skills: working memory and attention. Working memory is a cognitive function that allows one to temporarily hold and manipulate information for immediate tasks. Attention we measure is the cognitive process of selectively

focusing on specific information while ignoring other information. These two skills are widely regarded as fundamental pillars 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). To measure working memory, we developed a visual memory task based on (Luck and Vogel, 1997; Vogel et al., 2001; Oberauer, 2002; Halberda et al., 2006). In this task, students viewed a set of colored geometric shapes displayed on a screen for 10 seconds and were subsequently asked a question about either the color, position, or shape of the objects, without prior indication of which attribute would be asked. This design captures the ability to retain and manipulate visual information over a short period, a core component of working memory. The number of correct responses was standardized to have a mean of zero and a standard deviation of one in the control group, with higher values indicating stronger working memory.

We measured attention using the Flanker Task, a standard instrument in cognitive psychology (Eriksen and Eriksen, 1974). In this task, students are required to focus on a target stimulus while ignoring surrounding distracting stimuli (flankers). Performance reflects selective attention and impulse control—that is, the ability to attend to relevant information while suppressing interference from irrelevant inputs. Specifically, students were presented with a row of five arrows and were instructed to indicate the direction of the central arrow as quickly and accurately as possible, while ignoring the direction of the surrounding arrows. In congruent trials, the flankers point in the same direction as the target (e.g., <<<<<), whereas in incongruent trials, the flankers point in the opposite direction (e.g., <<><<). Successfully performing the task therefore requires suppressing misleading visual information in incongruent trials. We implemented both two-second and one-second response-time versions of the task and constructed an attention measure based on the difference in performance between incongruent and congruent trials. This measure was standardized to have a mean of zero and a standard deviation of one in the control group, with higher values indicating greater attentional capacity.

Both the working memory and flanker tests were administered in all waves, except for baseline, as shown in Online Appendix Table OA.0.1. A Full description of these two tests, including their visual structure, is given in the Online Appendix. Note that the intervention directly targeted working memory during its first year. Attention, by contrast, was not directly targeted in either year of the program. However, 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). We therefore conjecture that

the intervention improves both functions through their close cognitive interdependence.

4.1.2 Non-language-based Higher-order Skills

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

- **Abstract Reasoning:** We measured fluid intelligence with Raven's Progressive Matrices (Raven and Court, 1998), a widely used nonverbal test of abstract reasoning. Raven's provides a standardized measure of fluid intelligence, abstract reasoning and speed of learning. It is free from language ability and can be applied from age 5 onward.Through extensive exercises on pattern recognition and rotation, the intervention was expected to improve abstract reasoning and speed of learning, as intended to be measured by Raven's Progressive matrices.

- **Analytical Categorization:** We measured analytical categorization with the triad task (Kagan et al., 1963). Students are presented with three images and asked which two "go together." The choice reveals reasoning style: pairing car–bike indicates analytical reasoning, while bike–helmet indicates functional reasoning. This test allows us to capture students' emerging tendency toward abstract, categorical thought rather than context-dependent associations. Working on numerous categorization exercises using Venn diagrams, students are expected to develop a logical classification of objects.

- **Analogical Reasoning:** We assessed analogical reasoning through a task that required students to identify and transfer relationships across different contexts. Students were presented with structured multiple-choice questions based on relational mapping, where they had to infer how the connection between two items applied to a second pair. Unlike simple similarity judgments, the task emphasized recognizing underlying relational structures. The test included 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 were asked to select the correct answer from a set of options (e.g., milk, hay, sheep). Choosing "milk" indicates the ability to detect the relational logic (producer–product) and correctly extend it to a new context. This capacity to abstract relational patterns and generalize them beyond familiar examples is a key element of higher-order reasoning.

- **Logical reasoning:** This was a simple transitivity test in the form of if $A > B$ and $B > C$; A must be greater than C.

4.1.3 Language-based higher order thinking skills and Socio-cognition

Critical Thinking: We aimed to measure key aspects of critical thinking in children to assess whether the intervention improved the critical thinking process. Critical thinking refers to the ability to analyze, evaluate, and reason logically before forming a judgment or reaching a conclusion. It is a multidimensional construct, which makes it difficult to measure using a single tool. While survey-based instruments exist (e.g., (Alan and Mumcu, 2024); the critical thinking scale used in our curiosity paper), these are not suitable for the young age group targeted in this study. We therefore designed a task tailored to children who have a basic level of reading comprehension. The task aims to capture two key components of critical thinking: ability to evaluate evidence that is, inference by applying a logical reasoning, and logical justification of judgment. Below, we describe the design and classroom implementation of this task.

The task presents students with a short paragraph describing a lost item in a hypothetical school setting, alongside six possible locations, only one of which is correct. Students can earn ten gift tokens (convertible to visible prizes previously shown to them) if they identify the correct location, and zero otherwise. Each student is also given an endowment of six tokens, which they may use to purchase clues. Each clue is provided in a short paragraph and contains one piece of information that eliminates one of the six possible locations. Each clue costs one endowment token, and up to four clues are available. If a student purchases all four clues, the number of plausible locations is reduced to two, implying a 50% chance of selecting the correct location.

After reading the initial paragraph, students must decide whether to purchase a clue. If a student chooses not to purchase any clues, they are directed to select one of the six locations. After making their choice, they are asked why they decided not to purchase a clue and given a set of predefined response options as "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", "I am not curious about the correct answer".

If a student chooses to purchase a clue, they forgo one endowment token and are shown the first clue paragraph. This paragraph implicitly (quite subtly) indicates where the lost item cannot be. The student is then asked to identify which of the six locations is

ruled out based on the information provided. After the elimination of one option, they are asked to state what the clue was, capturing their ability to justify their reasoning. For this, they were asked to choose among three options regarding what the clue was, and the options were deliberately designed to be confusing to identify students' ability to justify their own reasoning regarding the elimination of the option. The student is then given the choice to purchase another clue or to stop and choose a location. This process repeats for up to four clues, depending on the student's choices. Figure OA.0.3 in Online Appendix depicts the decision sequences of the task.

Our measure of critical thinking is the ability to correctly eliminate the option and the correctly justifying the choice. However, note that these two indicators are only observable for students who choose to purchase a clue. If the treatment affects the likelihood of purchasing clues, conditioning on this decision would introduce selection bias. To address this, in the first round nearly all students were forced to purchase a clue. To avoid deception, students were informed that although they could express their preference, there was a very high probability they would be required to purchase a clue and a very small chance that their choice would be honored. This allowed us to observe evidence evaluation and justification ability for almost the entire sample in the first round. In subsequent rounds, students were free to choose whether or not to purchase clues. We estimate program impacts on (i) willingness to seek truth, measured by tokens spent on clues, (ii) ability to evaluate evidence and ability to justify reasoning.

Given the cognitive demands of the task and its reliance on reading comprehension, it was first implemented at the end of Grade 3 and repeated in Grade 4. We made some design changes in the second implementation of the task. Note first that given the payoff structure of the task, it is not payoff maximizing to purchase a clue. In the follow-up (Grade 4), we implemented the task by making the clue-taking free, thereby making the clue-taking payoff maximizing if the students are able to understand the clues and correctly eliminate the options. Another difference in the second year is that while asking whether to take a clue or not, we imposed clue taking in 90% of the cases. This design allows us to assess the ability of elimination and justification in all rounds. It also allows us to estimate the treatment effect on clue taking when clue taking is payoff maximizing.

Cognitive Empathy-Theory of Mind: While cognitive empathy does not map cleanly into either language-based or non-language-based cognition, we classify it as

a language-based higher-order thinking skill in this study. This classification reflects the fact that our measure—the Reading the Mind in the Eyes Test (RMET)—relies on reading comprehension and verbal processing. The test often involves understanding nuanced and abstract mental-state descriptions. Interpreting these descriptions and mapping them to visual cues requires some linguistic competence in addition to socio-cognitive inference. Prior research similarly emphasizes the language-dependent nature of RMET performance, particularly in child and adolescent populations (Peterson and Miller, 2012; Rosso and Riolfo, 2020).

4.2 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 Matrices and the Reading the Mind in the Eyes Test (RMET). At endline, we administered Raven's Progressive Matrices, the RMET, and an analogical reasoning task; these assessments were repeated at follow-up. When a classroom teacher changed during the study period, the newly assigned teacher was asked to complete the baseline survey and to take the Raven and RMET assessments. We then included them in our endline assessments.

In every wave, we also elicited a rich set of teacher beliefs, attitudes, and behaviors through structured surveys. These include measures of 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. In addition, we 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, while data collection was simultaneously conducted in their classrooms, ensuring that they were not present during student assessments. Example survey items for each construct are provided in the Online Appendix.

We hypothesize that by imposing sustained and substantial cognitive demands on teachers, the program may also improve teachers' higher-order reasoning skills. We examine this possibility using measures of abstract reasoning (Raven's Progressive Matrices), cognitive empathy (RMET), and analogical reasoning.

5 Data and Results

5.1 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	14803	88.20	88.11	0.36
Female (1,0)	14806	0.49	0.50	0.21
Refugee (1,0)	14806	0.08	0.07	0.48
Student cognitive skills				
Fluid IQ score	14806	0.00	0.08	0.21
Analytical categorization score	14806	0.00	-0.03	0.33
Cognitive empathy score	14665	0.00	0.05	0.33
Student academic outcomes				
Numeracy score	14478	-0.00	0.08	0.22
Literacy score	14038	-0.00	0.10	0.22

Because the program targets cognitive capacity and higher-order reasoning, we measure a wide set of cognitive, socio-cognitive, and higher-order outcomes. These outcomes are expected to be related. Table OA.0.2 shows the pairwise correlations among all measured child outcomes. Correlations are positive and statistically significant, but far from perfect. This suggests that the measures capture distinct dimensions of cognition. An important question is whether these skills predict academic performance. Table X shows that all measured skills are strongly predictive of both numeracy and literacy outcomes. Panel 2 presents the same relationships while controlling for baseline fluid IQ. The patterns remain similar. All skills retain predictive power for academic achievement. While the skills are correlated, each retains independent predictive power beyond the others. In the next section, we examine whether a program that aims to improve a subset of these skills can also improve other skills and, ultimately, academic

Table 2: Balance: Teachers Outcomes

	Count	Control Mean	Treatment Mean	P-value (Difference)
Class Characteristics				
Class Size	567	34.54	34.33	0.95
Class Refugee Share	567	0.11	0.11	0.73
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	562	-0.00	-0.01	0.94
Modern Teaching Methods	561	-0.00	0.05	0.59
Extrinsic Motivation	559	-0.00	0.04	0.46
Emotional Warmth	558	-0.00	-0.02	0.89
Teacher's Beliefs and Attitudes				
Growth-Mindset	560	0.00	-0.03	0.80
Professional Motivation	561	0.00	-0.14	0.19
Industriousness	561	-0.00	0.01	0.79
Self-Control	558	0.00	0.08	0.33
Sense of Responsibility	558	-0.00	0.11	0.18
Perceived Competence	562	-0.00	-0.03	0.83
Gender Bias	560	0.00	-0.07	0.48
Teacher's Personality Traits				
Critical-thinking	563	-0.00	0.18	0.06*
Curiosity	560	-0.00	-0.03	0.76
Conformism	558	0.00	-0.07	0.55
Teacher's Personal Well-being				
Mental Health	563	-0.00	0.06	0.50
Burnout	563	0.00	-0.03	0.75

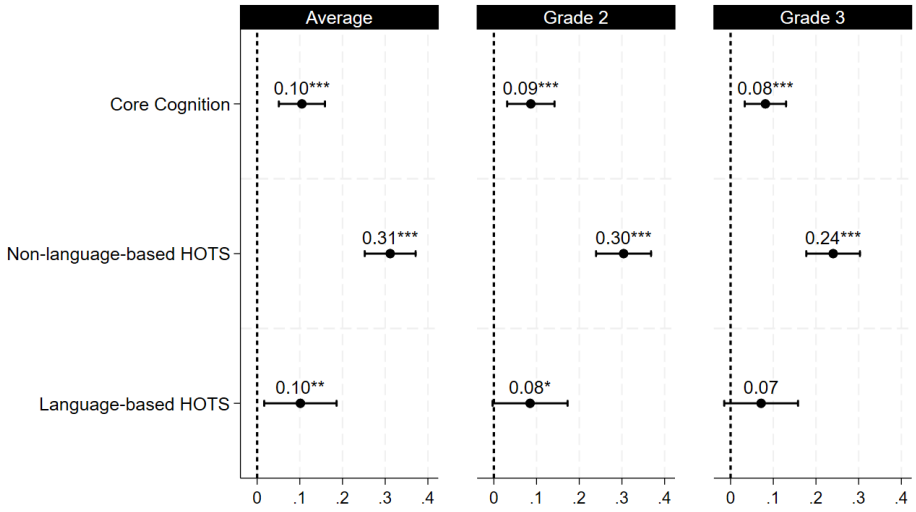
achievement.

5.2 Results

We first present the estimated treatment effects for three broad outcome categories: (i) core cognition (attention and working memory), (ii) non-language-based higher-order skills (analytical, abstract, analogical, and logical reasoning), and (iii) language-based higher-order skills (cognitive empathy, reading comprehension, and critical thinking). Column 1 of Figure 1 reports effects averaged over three years. Columns 2, 3, and 4 report effects separately by grade.

As shown in Figure 1, effects observed in Grade 2 persist over time, with no clear evidence that they either strengthen or fade. On average, we estimate an improvement of about 0.10 standard deviations in core cognition, 0.31 standard deviations in non-language-based higher-order skills, and 0.10 standard deviations in language-based higher-order skills. Overall, effects are weaker for language-based higher-order skills.

Figure 1: Treatment effects on aggregated cognitive outcomes: Grade 2, grade 3, and two-year average.



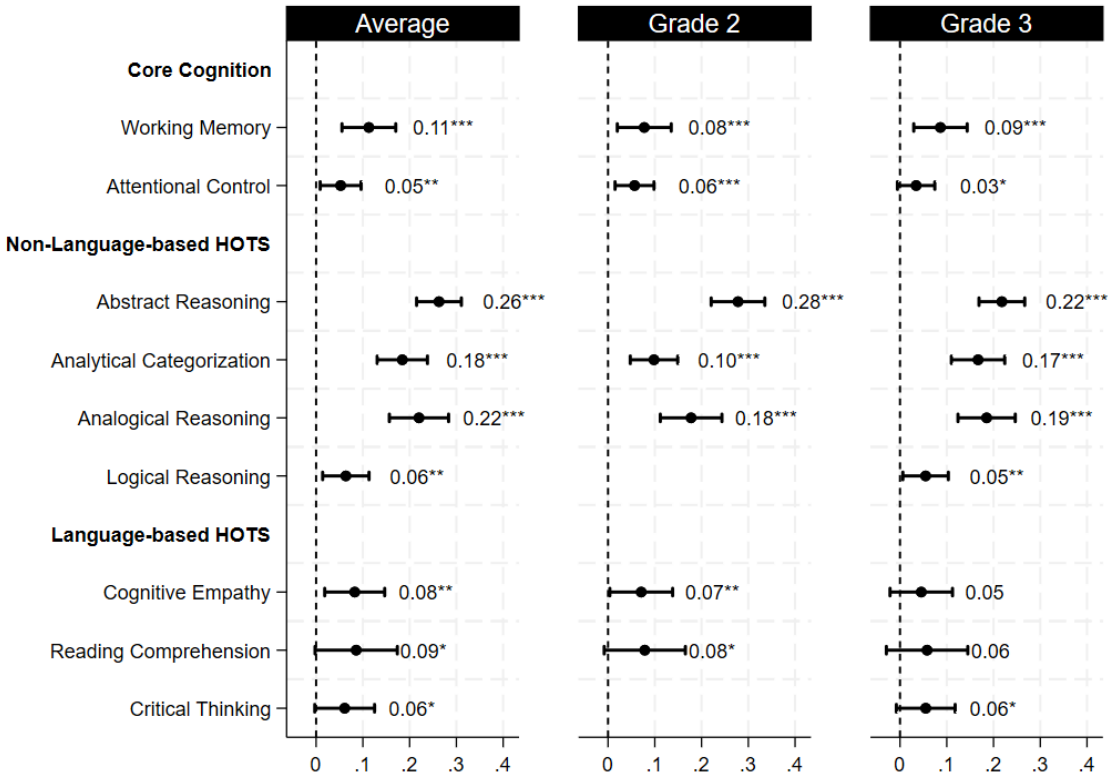
Notes: The figure plots the estimated effects of the treatment on the aggregated outcome indices. Points represent the estimated coefficients on the binary treatment indicator, and bars denote 95% confidence intervals. The aggregated outcomes are standardized using the control-group mean and standard deviation. Reported estimates are obtained from regressions controlling for 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.

Figure 2 also presents disaggregated results for each outcome category. The program improves all measured skills, with the strongest effects observed for non-language-based higher-order skills. We estimate sizable and persistent gains in abstract, analytical, and analogical reasoning, ranging from 0.18 to 0.26 standard deviations on average. In contrast, gains in language-based higher-order skills are more modest, though consistently positive across all measures. We find positive effects on cognitive empathy, reading comprehension, and critical thinking. While effect sizes are moderate, these results are notable given 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 core cognitive skills such as working memory and attention. Importantly, the results also point to malleability in higher-order reasoning skills at young ages. In the next section, we explore potential mechanisms underlying these effects.

To examine whether the program had differential effects across groups, we conduct a

Figure 2: Treatment effects on disaggregated cognitive outcomes: Grade 2, grade 3, and two-year average.



Notes: The figure plots the estimated treatment effects on the disaggregated outcome measures. Points represent the estimated coefficients on the binary treatment indicator, and bars denote 95% confidence intervals. The disaggregated outcomes are standardized using the control-group mean and standard deviation. Reported estimates are obtained from regressions controlling for 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.

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).[7] 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 seen in Figure 4, the program generates larger effects for girls across all three language-based measures. While point estimates for boys are positive, they are smaller and not statistically significant, and we cannot reject equality 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).[8]

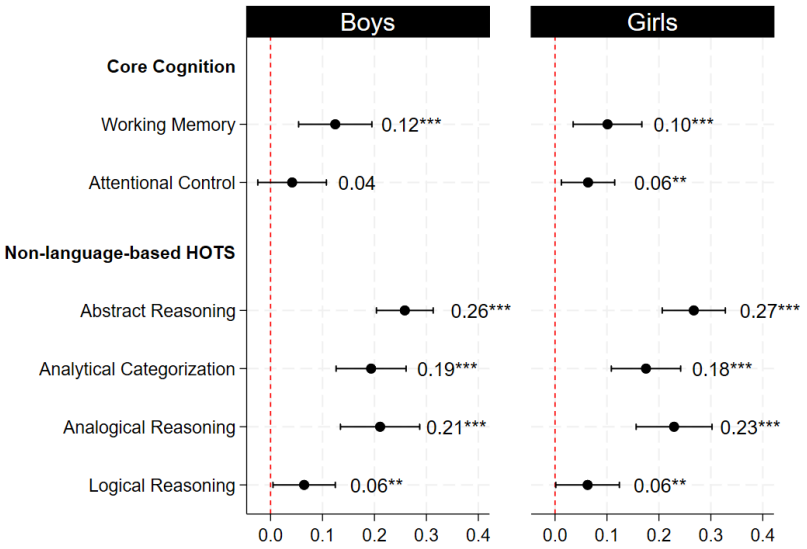
Figure 3 and 4 present heterogeneous treatment effects on non-language and language-based cognitive outcomes, respectively, by gender. We find no gender differences in core cognition or non-language-based higher-order skills. In contrast, girls exhibit clear advantages in language-based higher-order skills. The program is effective in improving language-based outcomes for girls, while effects on core cognition and non-language-based higher-order skills are positive and similar for both genders.

Figure 5 and 6 present heterogeneity by mother-tongue status, comparing Gaziantep and Şırnak. A striking pattern emerges. Non-language-based higher-order skills im-

[7]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).

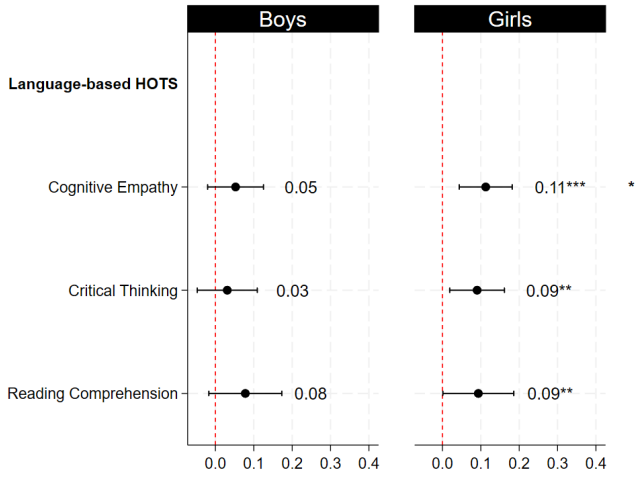
[8]Şırnak also includes a small number of non-Kurdish-speaking children, primarily children of civil servants and military personnel. Their number is too small to affect our results.

Figure 3: Effect on average core-cognition and non-language-based HOTS by gender



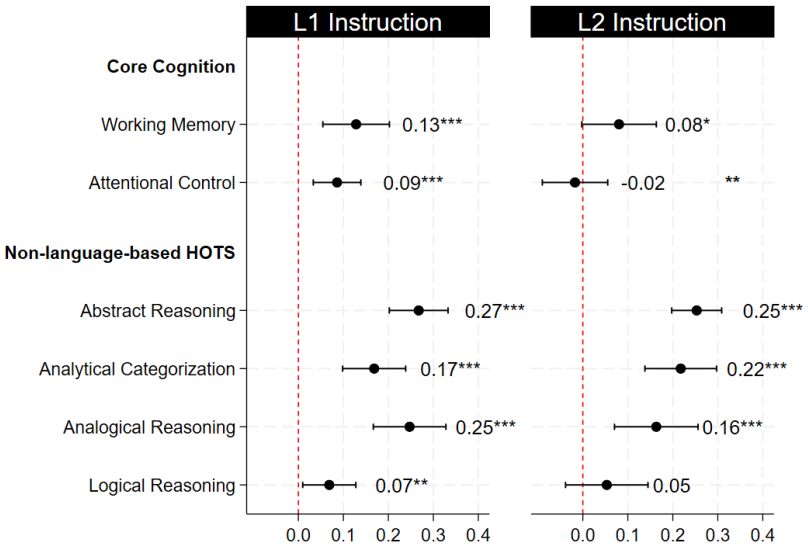
Notes: Points show estimated coefficients on the binary treatment indicator; bars denote 95% confidence intervals. The aggregated outcomes are standardized using the control-group mean and standard deviation. Each regression controls for a female indicator, a special-needs indicator, class size, a teacher-turnover indicator for each grade, and district fixed effects. Standard errors are clustered at the school level. (Stars only for the differential effects.)

Figure 4: Effect on average language-based HOTS by gender



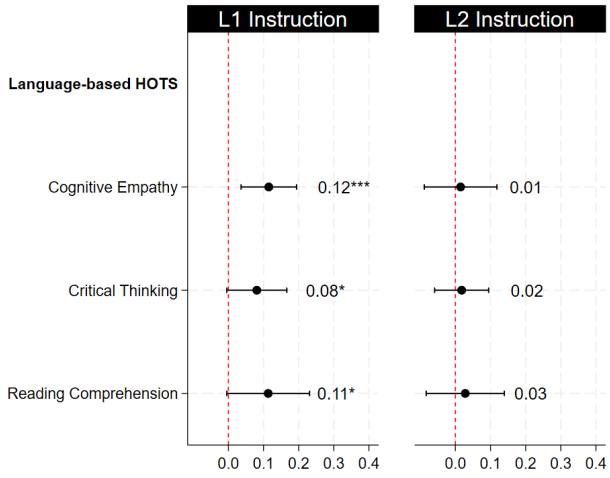
Notes: Points show estimated coefficients on the binary treatment indicator; bars denote 95% confidence intervals. The aggregated outcomes are standardized using the control-group mean and standard deviation. Each regression controls for a female indicator, a special-needs indicator, class size, a teacher-turnover indicator for each grade, and district fixed effects. Standard errors are clustered at the school level. (Stars only for the differential effects.)

Figure 5: Effect on average core-cognition and non-language-based HOTS by bilingualism



Notes: Points show estimated coefficients on the binary treatment indicator; bars denote 95% confidence intervals. The aggregated outcomes are standardized using the control-group mean and standard deviation. Each regression controls for a female indicator, a special-needs indicator, class size, a teacher-turnover indicator for each grade, and district fixed effects. Standard errors are clustered at the school level. (Stars only for the differential effects.)

Figure 6: Effect on average language-based HOTS by bilingualism

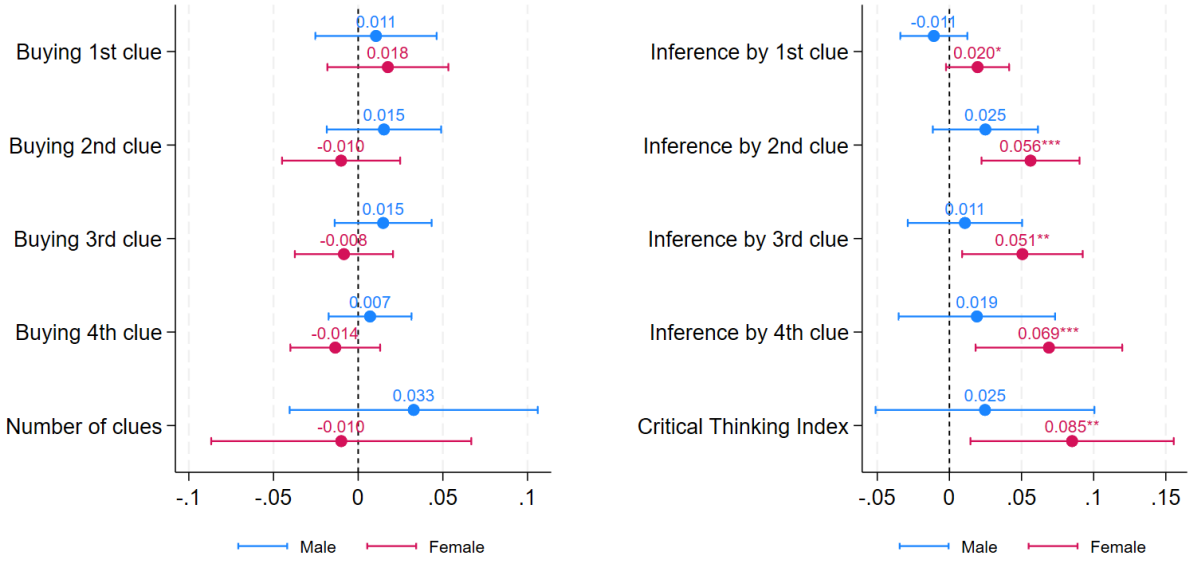


Notes: Points show estimated coefficients on the binary treatment indicator; bars denote 95% confidence intervals. The aggregated outcomes are standardized using the control-group mean and standard deviation. Each regression controls for a female indicator, a special-needs indicator, class size, a teacher-turnover indicator for each grade, and district fixed effects. Standard errors are clustered at the school level. (Stars only for the differential effects.)

prove similarly in both provinces. In contrast, we estimate no effects on language-based higher-order skills in the predominantly Kurdish-speaking province. Effects on core cognition are also weaker in Şırnak, though we observe some positive signs in Grade 3.

Differences become especially pronounced when we examine components of the critical thinking task. Panel 1 of Figure 7 reports treatment effects on the decision to purchase a clue in the Grade 3 task. We find no treatment effect and no gender heterogeneity for this behavior. Panel 2, however, reveals a clear gender difference in the critical thinking process itself. Girls perform significantly better in drawing inferences and justifying their reasoning. We observe a similar pattern by mother tongue in Figure 8. The program has a significantly greater impact on children's critical thinking when their mother tongue matches the language of instruction (Turkish).

Figure 7: Critical thinking game outcomes: Gender Heterogeneity



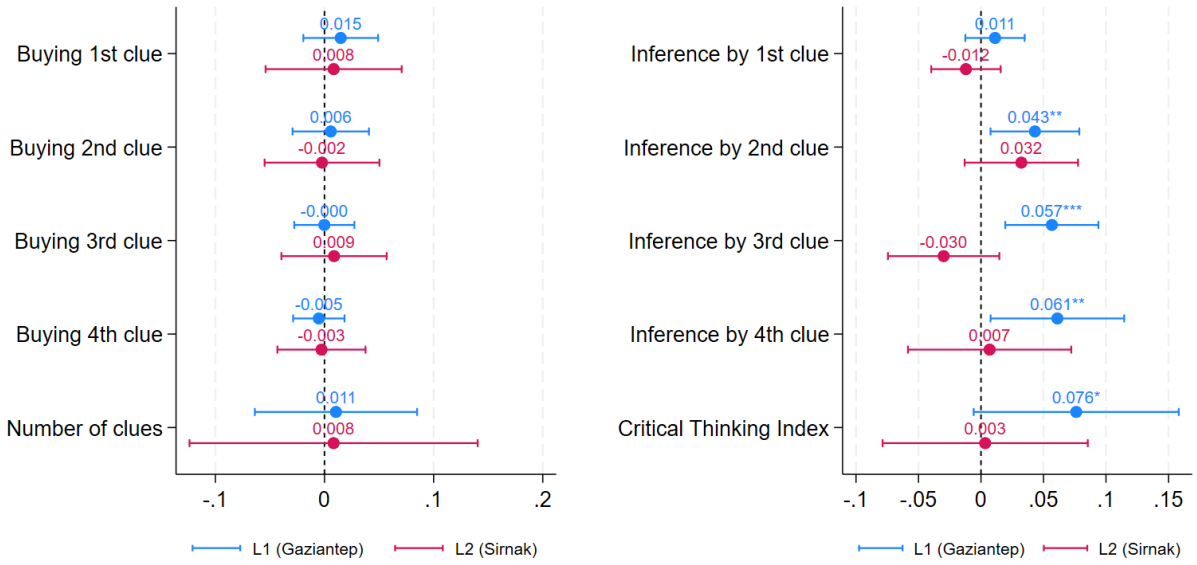
(a) Willingness to pay for truth

(b) Inference by the bought clue

Notes: Points show estimated coefficients on the binary treatment indicator; bars denote 95% confidence intervals. Outcomes are binary indicators unless otherwise noted; aggregated outcomes (e.g., the number of clues and the inference index) are standardized using the control group mean and standard deviation. Each regression controls for a female indicator, a special-needs indicator, class size, a teacher-turnover indicator for each grade, and district fixed effects. Standard errors are clustered at the school level.

Finally, we estimate the program's impact on teachers. Figure 10 presents the results. 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. In particular, they score higher on the Raven test (0.21 sd), the eye

Figure 8: Critical thinking game outcomes: Language Heterogeneity



(a) Willingness to pay for truth

(b) Inference by the bought clue

Notes: Points show estimated coefficients on the binary treatment indicator; bars denote 95% confidence intervals. Outcomes are binary indicators unless otherwise noted; aggregated outcomes (e.g., the number of clues and inference index) are standardized using the control-group mean and standard deviation. Each regression controls for a female indicator, a special-needs indicator, class size, a teacher-turnover indicator for each grade, and district fixed effects. Standard errors are clustered at the school level.

test (0.22 sd), and the analogical reasoning test (0.15 sd), with the latter significant at the 10% level. This improvement in teachers' cognitive skills may represent one pathway through which the program generates large effects on students: cognitively better-equipped teachers may become more effective at delivering instruction. We also find a substantial effect on teachers' beliefs. Treated teachers exhibit a 0.32 sd higher growth mindset, reflecting stronger beliefs in the role of effort in shaping intelligence. Finally, contrary to our initial expectations, treated teachers do not report higher burnout. Instead, they report 0.12 sd lower burnout. This suggests that teaching may become easier when students have stronger cognitive capacities.

6 Understanding the Results: Why was the Program Effective?

Cognitive development is understood as a dynamic, self-reinforcing system. Multiple cognitive processes evolve together by interacting with each other (Karmiloff-Smith, 1994; Welsh et al., 2010; Clements et al., 2016). Corroborating with this, we observe strong correlations among all cognitive skills we measured, as seen Table OA.0.2 in Appendix. In the early stages of childhood, core cognitive functions such as attention and working memory provide the necessary scaffolding for learning (Blair and Razza, 2007; Alloway and Alloway, 2010; Diamond, 2013; Blakey et al., 2020; Zelazo and Carlson, 2020; Doebel and Lillard, 2023). As children are exposed to structured learning environments (schools) and accumulate knowledge, they begin to develop higher-order reasoning skills — including abstraction, inference, categorization, and analogical thinking — which allow them to recognize patterns, form generalizations, and apply knowledge across contexts.

An intervention that can improve higher order reasoning might enable more effective learning; better learning experiences further strengthen higher-order reasoning, deepening critical thinking. The idea of the intervention was that one might accelerate and strengthen this development by pushing cognitive limits of children, as suggested by the premise of neural plasticity (Greenough et al., 1987; Knudsen et al., 2006; Diamond et al., 2007; Kolb and Gibb, 2011; Distefano et al., 2020; Slattery et al., 2022) What did we expect from the program? To accelerate the development of these skills and improve learning, despite crowding out. However, the direction of the causality is difficult to establish. Cognitive science suggests that general cognitive functions such as attention

Figure 9: Treatment effect on teaching styles and teachers' cognitive outcomes

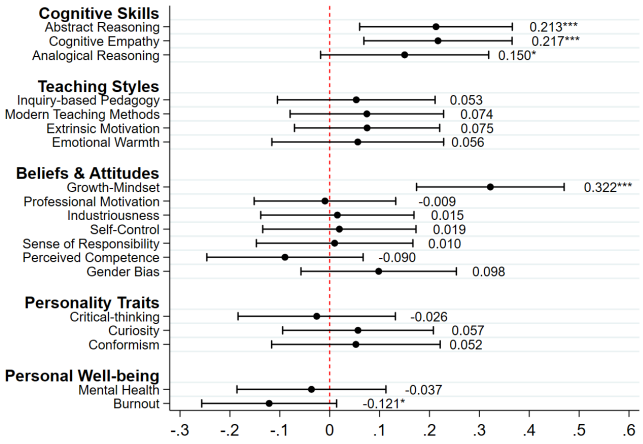
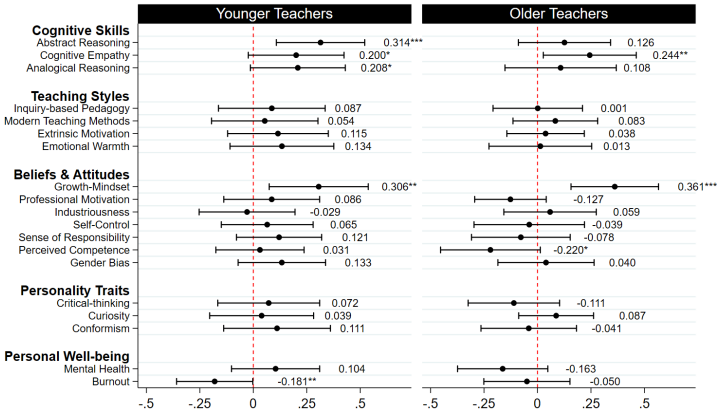


Figure 10: Heterogeneous effects on teaching styles and teachers' cognitive outcomes by age



and working memory provide the initial scaffolding for learning, while higher-order reasoning gradually emerges through interaction with learning experiences, in particular, knowledge accumulation. Over time, these processes become mutually reinforcing, generating a dynamic feedback loop in which improved reasoning reshapes learning strategies and further strengthens core cognitive functions. The purpose of the program was pushing cognitive limits of kids who dot get to be pushed enough. Not to establish the direction of processes in the brain.

A Dynamic Human Capital Accumulation Model of HOTS

Let i index children and $t \in \{0, 1, 2, 3, 4\}$ denote annual survey waves. Wave $t = 0$ is a pre-treatment baseline collected before any exposure to the intervention. Waves

$t = 1, 2, 3, 4$ are post-baseline waves. The program starts shortly after the baseline and runs for two school years, affecting the transitions from $t = 0$ to $t = 1$ and from $t = 1$ to $t = 2$.

Let $D_i \in \{0, 1\}$ indicate random assignment to the program. Conditional on sampling, assignment is independent of baseline skills and covariates:

$$D_i \perp\!\!\!\perp \boldsymbol{\theta}_{i0}, \mathbf{X}_i,$$

where $\boldsymbol{\theta}_{i0}$ denotes baseline latent skills and $\mathbf{X}_i$ denotes observed child characteristics (including district fixed effects).

Latent Skills and Measurement System

We assume that at each wave t child i is endowed with a vector of latent skills

$$\boldsymbol{\theta}_{it} = \begin{pmatrix} \theta_{it}^A \\ \theta_{it}^M \\ \theta_{it}^H \\ \theta_{it}^Y \end{pmatrix},$$

where θ_{it}^A is attention, θ_{it}^M is working memory, θ_{it}^H is higher-order thinking skills, and θ_{it}^Y is academic learning.

For each domain $d \in \{A, M, H, Y\}$ and test $k = 1, \ldots, K_d$, we observe a test score $S_{it}^{d,k}$ that is a noisy indicator of the corresponding latent skill:

$$S_{it}^{d,k} = \mu_t^{d,k} + \lambda^{d,k} \theta_{it}^d + \nu_{it}^{d,k}, \tag{1}$$

where $\mu_t^{d,k}$ is a time-specific intercept, $\lambda^{d,k}$ is a factor loading, and $\nu_{it}^{d,k}$ is a mean-zero measurement error, independent of $\boldsymbol{\theta}_{it}$ conditional on child characteristics. To identification for each domain d, we normalize one loading to one and fix the mean and variance of inital conditions θ_{i0}^d.

Stacking all scores for child i at time t in the vector $\mathbf{S}_{it}$ and all loadings in the matrix $\boldsymbol{\Lambda}$ gives the compact measurement equation

$$\mathbf{S}_{it} = \boldsymbol{\mu}_t + \boldsymbol{\Lambda} \boldsymbol{\theta}_{it} + \boldsymbol{\nu}_{it}, \tag{2}$$

with $\boldsymbol{\nu}_{it}$ collecting all measurement errors.

Dynamic Skill Formation

Attention and working memory are closely related components of executive function. The cognitive literature does not establish a clear causal ordering between them. Rather than imposing a directional restriction, we allow for dynamic interdependence between the two domains while refraining from a structural interpretation of these links.

Specifically, we adopt the following transition matrix for the latent skill vector $\boldsymbol{\theta}_{it} = (\theta_{it}^A, \theta_{it}^M, \theta_{it}^Y, \theta_{it}^H)'$:

$$\mathbf{A} = \begin{pmatrix} a_{AA} & a_{AM} & 0 & a_{AH} \\ a_{MA} & a_{MM} & 0 & a_{MH} \\ a_{YA} & a_{YM} & a_{YY} & a_{YH} \\ 0 & 0 & a_{HY} & a_{HH} \end{pmatrix}. \tag{3}$$

The diagonal elements capture persistence within each domain. The off-diagonal coefficients a_{YA} and a_{YM} reflect how attention and memory contribute to subsequent learning, a_{HY} captures how learning fosters higher-order thinking, and a_{AH}, a_{MH}, a_{YH} capture feedback effects from higher-order skills to core cognition and learning.

The coefficients a_{AM} and a_{MA} represent dynamic interactions between attention and working memory. Because theory does not provide a clear causal direction between these two domains, we interpret these parameters as reduced-form measures of interdependence within a broader executive control system rather than as structural causal effects.

If we go with this Matrix the component-wise equations:

Writing (**??**) component-wise under the transition structure

$$\boldsymbol{\theta}_{it} = (\theta_{it}^A, \theta_{it}^M, \theta_{it}^Y, \theta_{it}^H)', \qquad \mathbf{A} = \begin{pmatrix} a_{AA} & a_{AM} & 0 & a_{AH} \\ a_{MA} & a_{MM} & 0 & a_{MH} \\ a_{YA} & a_{YM} & a_{YY} & a_{YH} \\ 0 & 0 & a_{HY} & a_{HH} \end{pmatrix},$$

we obtain:

$$\theta^A_{i,t+1} = \alpha_A + a_{AA}\theta^A_{it} + a_{AM}\theta^M_{it} + a_{AH}\theta^H_{it} + \mathbf{b}'_A\mathbf{X}_i + \gamma^A_P D_i \mathbb{1}\{t \in \{0,1\}\} + \gamma^A_F D_i \mathbb{1}\{t \in \{2,3\}\} + \varepsilon^A_{i,t+1}, \tag{4}$$

$$\theta^M_{i,t+1} = \alpha_M + a_{MA}\theta^A_{it} + a_{MM}\theta^M_{it} + a_{MH}\theta^H_{it} + \mathbf{b}'_M\mathbf{X}_i + \gamma^M_P D_i \mathbb{1}\{t \in \{0,1\}\} + \gamma^M_F D_i \mathbb{1}\{t \in \{2,3\}\} + \varepsilon^M_{i,t+1} \tag{5}$$

$$\theta^Y_{i,t+1} = \alpha_Y + a_{YA}\theta^A_{it} + a_{YM}\theta^M_{it} + a_{YY}\theta^Y_{it} + a_{YH}\theta^H_{it} + \mathbf{b}'_Y\mathbf{X}_i + \gamma^Y_P D_i \mathbb{1}\{t \in \{0,1\}\} + \gamma^Y_F D_i \mathbb{1}\{t \in \{2,3\}\} \tag{6}$$

$$\theta^H_{i,t+1} = \alpha_H + a_{HY}\theta^Y_{it} + a_{HH}\theta^H_{it} + \mathbf{b}'_H\mathbf{X}_i + \gamma^H_P D_i \mathbb{1}\{t \in \{0,1\}\} + \gamma^H_F D_i \mathbb{1}\{t \in \{2,3\}\} + \varepsilon^H_{i,t+1}. \tag{7}$$

The diagonal coefficients a_{dd} capture persistence within each domain. The coefficients (a_{YA}, a_{YM}) capture how core cognition predicts future learning, a_{HY} captures how learning predicts future higher-order thinking, and (a_{AH}, a_{MH}, a_{YH}) capture feedback from higher-order thinking into core cognition and learning. The coefficients (a_{AM}, a_{MA}) allow for dynamic interdependence between attention and working memory without imposing a causal ordering between the two.

Interpretation as a Skill Formation Technology

The linear system in (**??**) can be interpreted as a local log-linear approximation to a more general nonlinear technology of skill formation in the spirit of Heckman and co-authors. For example, for each domain d one can write

$$\theta^d_{i,t+1} = f_d\Big(\boldsymbol{\theta}_{it}, I_{it}, \mathbf{X}_i, D_i, \varepsilon^d_{i,t+1}\Big),$$

where I_{it} denotes (possibly unobserved) investments and f_d embeds self-productivity, cross-skill productivity, and dynamic complementarity between existing skills and investments. The linear specification (**??**) then provides an estimable representation that captures these features through the elements of $\mathbf{A}$ and the treatment shifters $\boldsymbol{\Gamma}_P$ and $\boldsymbol{\Gamma}_F$.

State-Space Representation

The combination of the dynamic equations and the measurement system defines a state-space model. The *state equation* is given by

$$\boldsymbol{\theta}_{i,t+1} = \boldsymbol{\alpha} + \mathbf{A}\boldsymbol{\theta}_{it} + \mathbf{B}\mathbf{X}_i + \boldsymbol{\Gamma}_P D_i \mathbb{1}\{t \in \{0,1\}\} + \boldsymbol{\Gamma}_F D_i \mathbb{1}\{t \in \{2,3\}\} + \boldsymbol{\varepsilon}_{i,t+1}, \tag{8}$$

with initial condition (**??**). The *measurement equation* is (2):

$$\mathbf{S}_{it} = \boldsymbol{\mu}_t + \boldsymbol{\Lambda}\boldsymbol{\theta}_{it} + \boldsymbol{\nu}_{it}. \tag{9}$$

We assume $\boldsymbol{\varepsilon}_{i,t}$ and $\boldsymbol{\nu}_{it}$ are jointly Gaussian, independent across t, and independent of D_i and $\mathbf{X}_i$.

Identification

Random assignment of D_i ensures that, in expectation, treated and control children have the same distribution of baseline skills and covariates. The control group dynamics from $t = 0$ to $t = 4$ identify the elements of $\mathbf{A}$ and the variance-covariance matrices $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_\varepsilon$ (up to the usual dynamic panel considerations). Differences in the evolution of treated and control children across the program and follow-up transitions identify the treatment parameters $\boldsymbol{\Gamma}_P$ and $\boldsymbol{\Gamma}_F$, by skill domain.

The factor structure in the measurement equation, together with the normalizations on loadings and baseline distributions, identifies the scale and location of each latent skill and the loading matrix $\boldsymbol{\Lambda}$. Having multiple waves and multiple tests per domain is crucial for separating true skill dynamics from measurement error.

The triangular structure of $\mathbf{A}$ delivers a recursive ordering within each period, allowing us to interpret cross-skill coefficients as capturing the direction of influence from core cognition to higher-order skills and academic achievement. Circular causality then appears through the multi-period propagation of shocks across domains via $\mathbf{A}^2, \mathbf{A}^3$, and so on.

Estimation

Under the Gaussian assumptions on $(\boldsymbol{\varepsilon}_{i,t}, \boldsymbol{\nu}_{it}, \mathbf{u}_i)$, the model can be estimated by maximum likelihood using the Kalman filter and smoother. The likelihood integrates over the latent skill paths $\{\boldsymbol{\theta}_{it}\}_{t=0}^{4}$ implied by the state equation (8) and the initial condition (**??**), and the measurement equation (9). The parameter vector

$$\Psi = \Big(\boldsymbol{\alpha}, \mathbf{A}, \mathbf{B}, \boldsymbol{\Gamma}_P, \boldsymbol{\Gamma}_F, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_4, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_\varepsilon, \boldsymbol{\Sigma}_\nu\Big)$$

is chosen to maximize the joint log-likelihood of the observed test scores $\{\mathbf{S}_{it}\}$ given $(D_i, \mathbf{X}_i)$.

As a robustness check, we can also implement Bayesian estimation, placing priors on Ψ and drawing posterior distributions using simulation-based methods. The repeated

measurement of multiple skill domains over five waves, combined with the experimental variation from the two-year randomized program, provides strong leverage to identify the dynamic skill formation process and the way improvements in one domain propagate to others over time. (NO NEED TO DO THIS).

Post-Estimation Analysis

After estimating the parameter vector $\widehat{\Psi}$ by maximum likelihood, the system is fully parameterized. In this section we describe how we use the estimated model to recover latent skills, study the dynamics of skill formation, decompose the intervention effects, and conduct policy-relevant counterfactuals. The analysis closely follows the structural literature on skill formation (e.g., Cunha and Heckman, 2010; Attanasio et al., 2020).

Recovering Latent Skill Trajectories:

Given $\widehat{\Psi}$, the Kalman smoother yields the smoothed posterior means of latent skills:

$$\widehat{\boldsymbol{\theta}}_{it} = \mathbb{E}\Big(\boldsymbol{\theta}_{it} \,\Big|\, \mathbf{S}_{i0}, \ldots, \mathbf{S}_{i4};\, \widehat{\Psi}\Big)\,.$$

These estimates remove measurement error and produce de-noised trajectories of attention, memory, higher-order thinking, and academic skills for each child. We use them to describe:

- baseline skill distributions,
- the evolution of each domain over the five waves.

The smoothed trajectories also provide a direct check on whether the structural model replicates the reduced-form patterns observed in the data.

Counterfactual Simulations

Once the model is estimated, we can simulate the evolution of skills under alternative policies. Let $\{\boldsymbol{\theta}^*_{it}\}$ denote counterfactual paths. For each child, we simulate:

(i) No-treatment counterfactual

$$\boldsymbol{\theta}^*_{i,t+1} = \widehat{\boldsymbol{\alpha}} + \widehat{\mathbf{A}}\,\boldsymbol{\theta}^*_{it} + \widehat{\mathbf{B}}\,\mathbf{X}_i + \widehat{\boldsymbol{\varepsilon}}_{i,t+1},$$

corresponding to $D_i = 0$.

(ii) Full treatment counterfactual Same system with $D_i = 1$. The difference between (i) and (ii) delivers model-based treatment effects.

7 Conclusion

In this paper, we study whether formal schooling can improve the quality of children's thinking. We design and evaluate an intensive cognitive expansion program implemented in primary schools serving socioeconomically disadvantaged students. The program deliberately challenges students and teachers with cognitively demanding activities that target core cognitive capacities and higher-order reasoning skills. We implement the program over two academic years in 132 primary schools and follow more than 14,500 students for three years in Turkey.

We measure a broad set of outcomes capturing core cognition, academic skills, socio-cognitive abilities, and higher-order reasoning. We also introduce a novel language-based task that measures key components of critical thinking. We find that sustained exposure to cognitively demanding instructional activities improves executive functions such as working memory and attention. The program also strengthens non-language-based higher-order reasoning skills, including abstract, analytical, analogical, and logical reasoning. In contrast, improvements in language-based higher-order reasoning appear primarily among girls. Treated girls display stronger critical thinking processes, higher verbal reasoning, and greater cognitive empathy relative to girls in control schools. We also find that the program improves teachers' reasoning skills and increases the likelihood that they adopt a growth mindset about students' cognitive potential.

Two caveats deserve attention. First, we do not observe students long enough to determine how persistent the gains in cognition and reasoning will be. However, because skill formation follows a cumulative process, improvements in foundational cognitive capacities may generate lasting benefits by facilitating future learning. Second, implementing a cognitively demanding program as a stand-alone intervention may prove challenging for teachers. Integrating similar activities directly into the official curriculum may therefore represent a more practical approach. Such integration would likely ease implementation and further reduce program costs, which are already relatively modest.

Overall, our findings suggest that schools can do more than transmit academic knowl-

edge. With the right instructional approach, they can strengthen the cognitive processes that shape how children think, learn, and reason. In a world that increasingly demands critical and rational thinkers, improving the quality of children's thinking should become a central objective of education systems. Our study provides a first step in that direction by showing that schools can deliberately cultivate these capacities through sustained cognitive challenge. By mobilizing classrooms to engage students in deeper reasoning and structured problem solving, education systems may be able to foster stronger foundations for learning, innovation, and long-run human progress.

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.

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.

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.

Clancy Blair and Rachel Peters Razza. Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child development*, 78(2):647–663, 2007.

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. Executive functions. *Annual review of psychology*, 64(1):135–168, 2013.

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.

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.

C Kirabo Jackson. What do test scores miss? the importance of teacher effects on non–test score outcomes. *Journal of political Economy*, 126(5):2072–2107, 2018.

J Kagan, HA Moss, and IE Siegel. The psychological significance of styles of categorization. *Basic cognitive processes in children. Monographs of the Society for Research in Child Development*, 28(2), 1963.

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.

Online Appendix: Not for Publication

Figure OA.0.1: Study provinces in Turkey



Notes: The figure highlights the two provinces included in the study sample: Gaziantep and Şırnak.

Figure OA.0.2: Timeline of program implementation and data collection

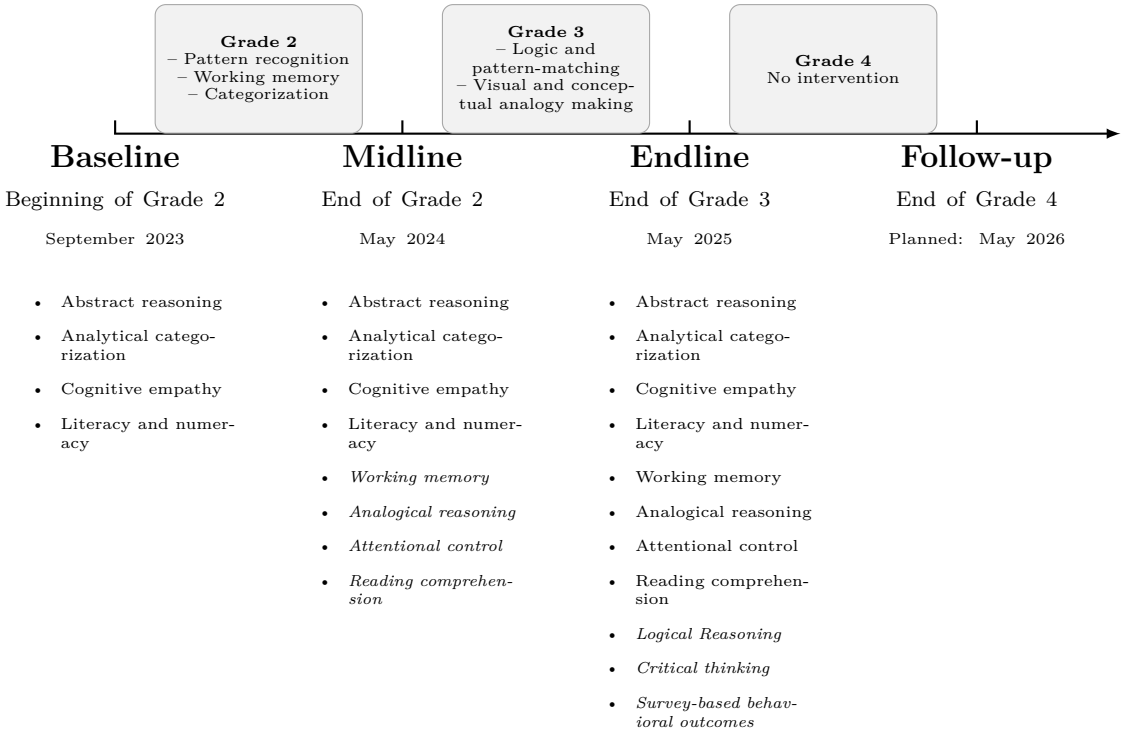


Table OA.0.1: Timing of Student Measurements

Measure	**Baseline**	**Midline**	**Endline**	**Follow-up**
Core cognition				
Working memory		✓	✓	❍
Attentional control		✓	✓	❍
Non-language based higher-order thinking skills				
Abstract reasoning	✓	✓	✓	❍
Analytical categorization	✓	✓	✓	❍
Analogical reasoning		✓	✓	❍
Logical reasoning			✓	❍
Language based higher-order thinking skills				
Critical thinking			✓	❍
Cognitive empathy	✓	✓	✓	❍
Reading comprehension		✓	✓	❍
Academic outcomes				
Literacy	✓	✓	✓	❍
Numeracy	✓	✓	✓	❍
Other outcomes				
Survey-based behavioral outcomes			✓	❍

Notes: ✓ = collected; ❍ = planned (in the follow up). Baseline is administered at the beginning of Grade 2; midline at the end of Grade 2; endline at the end of Grade 3; and follow-up at the end of Grade 4.

Table OA.0.2: Correlations Across Outcomes in the Control Group

	Working Memory	Attentional Control	Abstract Reasoning	Analytical Reasoning	Analogical Reasoning	Logical Reasoning	Cognitive Empathy	Critical Thinking	Reading Comprehension
Working Memory	1.00								
Attentional Control	0.11***	1.00							
Abstract Reasoning	0.26***	0.23***	1.00						
Analytical Reasoning	0.04**	0.03**	0.05**	1.00					
Analogical Reasoning	0.23***	0.20***	0.53***	0.12***	1.00				
Logical Reasoning	0.13***	0.13***	0.32***	0.03***	0.29***	1.00			
Cognitive Empathy	0.19***	0.17***	0.44***	0.02	0.38***	0.32***	1.00		
Critical Thinking	0.12***	0.10***	0.25***	0.03**	0.25***	0.16***	0.20***	1.00	
Reading Comprehension	0.24***	0.22***	0.54***	0.10***	0.49***	0.39***	0.47***	0.28***	1.00

Figure OA.0.3: Decision Tree for the Critical Thinking Task

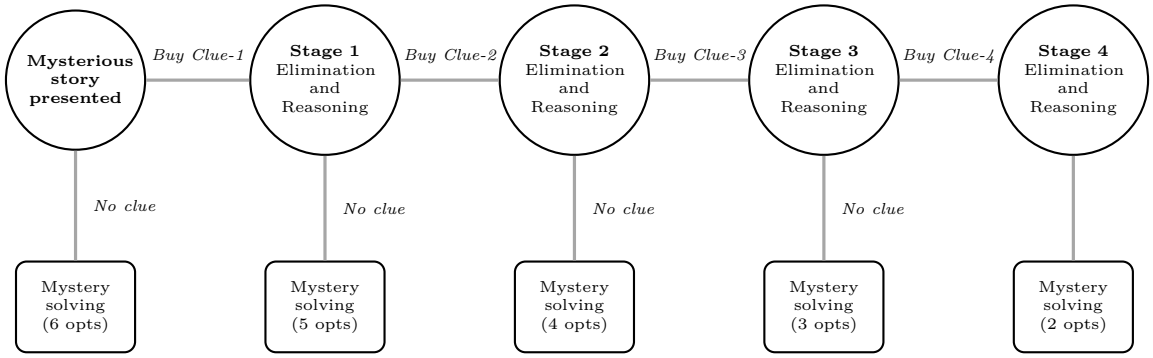


Table OA.0.3: Balance of student outcomes in control and treatment group: By gender and mother tongue

	Count	Control Mean	Treatment Mean	P-value(Difference)
Panel A: Female students				
Student demographics				
Age in months	7346	87.88	87.90	0.94
Refugee (1,0)	7347	0.08	0.07	0.69
Student cognitive skills				
Abstract reasoning	7347	0.02	0.09	0.22
Analytical categorization	7347	-0.07	-0.09	0.60
Cognitive empathy	7270	0.05	0.10	0.36
Student academic outcomes				
Numeracy score	7146	-0.08	0.02	0.18
Literacy score	6925	0.01	0.12	0.19
Panel B: Male students				
Student demographics				
Age in months	7457	88.51	88.33	0.15
Refugee (1,0)	7459	0.08	0.07	0.32
Student cognitive skills				
Abstract reasoning	7459	-0.02	0.06	0.24
Analytical categorization	7459	0.07	0.03	0.23
Cognitive empathy	7395	-0.04	0.00	0.35
Student academic outcomes				
Numeracy score	7332	0.08	0.15	0.32
Literacy score	7113	-0.01	0.08	0.29
Panel C: Students in Gaziantep				
Student demographics				
Age in months	9773	88.62	88.55	0.50
Female (1,0)	9776	0.49	0.50	0.34
Refugee (1,0)	9776	0.11	0.10	0.45
Student cognitive skills				
Abstract reasoning	9776	0.03	0.16	0.13
Analytical categorization	9776	-0.10	-0.16	0.11
Cognitive empathy	9658	0.02	0.12	0.15
Student academic outcomes				
Numeracy score	9596	-0.01	0.11	0.14
Literacy score	9316	-0.02	0.12	0.14
Panel D: Students in Sirnak				
Student demographics				
Age in months	5030	87.35	87.29	0.52
Female (1,0)	5030	0.49	0.50	0.42
Refugee (1,0)	5030	0.00	0.00	0.32
Student cognitive skills				
Abstract reasoning	5030	-0.05	-0.07	0.71
Analytical categorization	5030	0.21	0.21	0.80
Cognitive empathy	5007	-0.04	-0.08	0.45
Student academic outcomes				
Numeracy score	4882	0.02	0.03	0.85
Literacy score	4722	0.05	0.05	0.83

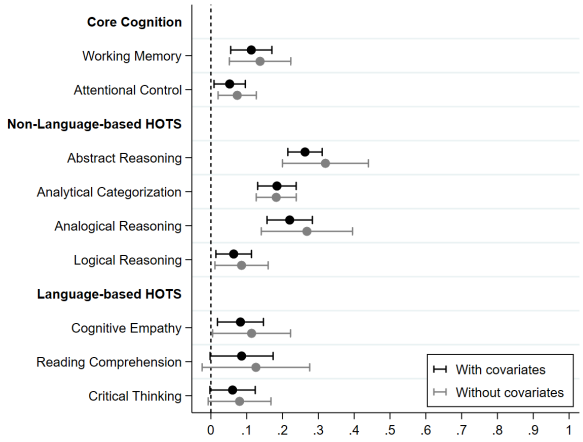
Table OA.0.4: Balance of teacher outcomes in control and treatment group: L1 and L2 Instruction

	Count	Control Mean	Treatment Mean	P-value of Difference
Panel A: Teachers and Classes in Gaziantep				
Class Characteristics				
Class Size	367	37.48	37.26	0.98
Class Refugee Share	367	0.16	0.17	0.70
Teacher Demographics				
Age	367	36.32	37.88	0.29
Male (1,0)	367	0.34	0.32	0.71
Married (1,0)	366	0.62	0.72	0.11
Teacher Cognitive Skills				
Fluid IQ	367	-0.13	-0.01	0.36
Cognitive Empathy	367	0.01	-0.08	0.47
Teaching Styles				
Inquiry-based Pedagogy	365	0.02	0.12	0.35
Modern Teaching Methods	364	-0.06	0.11	0.15
Extrinsic Motivation	363	0.02	0.07	0.52
Emotional Warmth	362	0.02	0.08	0.57
Teacher's Beliefs and Attitudes				
Growth-Mindset	364	-0.13	-0.02	0.19
Professional Motivation	364	-0.04	-0.13	0.54
Industriousness	364	0.04	0.11	0.52
Self-Control	364	0.01	0.07	0.62
Sense of Responsibility	364	-0.04	0.17	0.03**
Perceived Competence	365	0.08	0.09	0.88
Gender Bias	364	0.07	-0.06	0.25
Teacher's Personality Traits				
Critical-thinking	366	0.04	0.26	0.07*
Curiosity	364	-0.00	0.03	0.73
Conformism	364	0.10	0.01	0.36
Teacher's Personal Well-being				
Mental Health	366	0.01	0.14	0.26
Burnout	366	0.06	0.02	0.69
Panel B: Teachers and Classes in Sirnak				
Class Characteristics				
Class Size	200	28.89	29.20	0.84
Class Refugee Share	200	0.00	0.01	0.48
Teacher Demographics				
Age	200	29.62	29.97	0.64
Male (1,0)	200	0.34	0.38	0.64
Married (1,0)	199	0.44	0.40	0.64
Teacher Cognitive Skills				
Fluid IQ	199	0.25	0.08	0.11
Cognitive Empathy	197	-0.02	-0.05	0.84
Teaching Styles				
Inquiry-based Pedagogy	197	-0.04	-0.25	0.16
Modern Teaching Methods	197	0.12	-0.06	0.24
Extrinsic Motivation	196	-0.03	-0.00	0.72
Emotional Warmth	196	-0.03	-0.19	0.29
Teacher's Beliefs and Attitudes				
Growth-Mindset	196	0.25	-0.05	0.07*
Professional Motivation	197	0.07	-0.17	0.16
Industriousness	197	-0.07	-0.16	0.66
Self-Control	194	-0.01	0.11	0.34
Sense of Responsibility	194	0.09	0.01	0.56
Perceived Competence	197	-0.15	-0.24	0.52
Gender Bias	196	-0.14	-0.09	0.67
Teacher's Personality Traits				
Critical-thinking	197	-0.09	0.02	0.55
Curiosity	196	0.01	-0.13	0.31
Conformism	194	-0.20	-0.22	0.98
Teacher's Personal Well-being				
Mental Health	197	-0.03	-0.10	0.62
Burnout	197	-0.13	-0.12	1.00

Table OA.0.5: Correction for Multiple Hypothesis Testing

	Original	**Romano–Wolf**
Panel A: Full Sample		
Core Cognition	0.0002	0.0020
Non-language-based HOTS	0.0000	0.0020
Language-based HOTS	0.0195	0.0060
Panel B: Boys		
Core Cognition	0.0040	0.0020
Non-language-based HOTS	0.0000	0.0020
Language-based HOTS	0.1204	0.0539
Panel C: Girls		
Core Cognition	0.0008	0.0040
Non-language-based HOTS	0.0000	0.0020
Language-based HOTS	0.0043	0.0040
Panel D: L1 instruction		
Core Cognition	0.0001	0.0020
Non-language-based HOTS	0.0000	0.0020
Language-based HOTS	0.0141	0.0060
Panel E: L2 instruction		
Core Cognition	0.2678	0.3972
Non-language-based HOTS	0.0000	0.0020
Language-based HOTS	0.6932	0.6547

Figure OA.0.4: Robustness check: Controlling for covariates



Notes: Points show estimated coefficients on the binary treatment indicator; bars denote 95% confidence intervals. The aggregated outcomes are standardized using the control-group mean and standard deviation. Each regression controls for a female indicator, a special-needs indicator, class size, a teacher-turnover indicator for each grade, and district fixed effects. Standard errors are clustered at the school level.

OA.1 Testing Instruments and Instructions

OA.1.1 Student Instruments

Figure OA.1.1: Sample screen in the working memory task

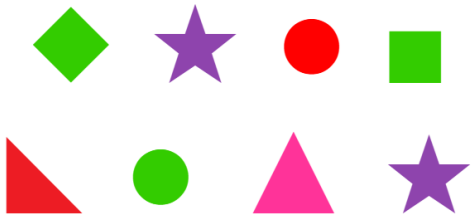
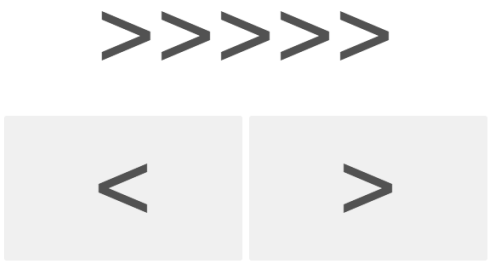
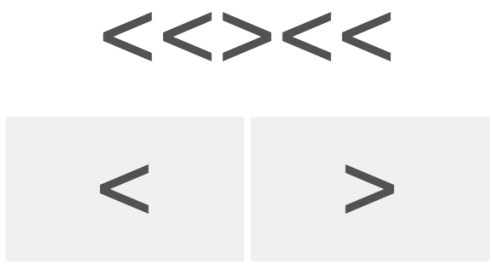


Figure OA.1.2: Sample screens in the flanker task



(a) Congruent trial



(b) Incongruent trial

Figure OA.1.3: Sample screen in the abstract reasoning test

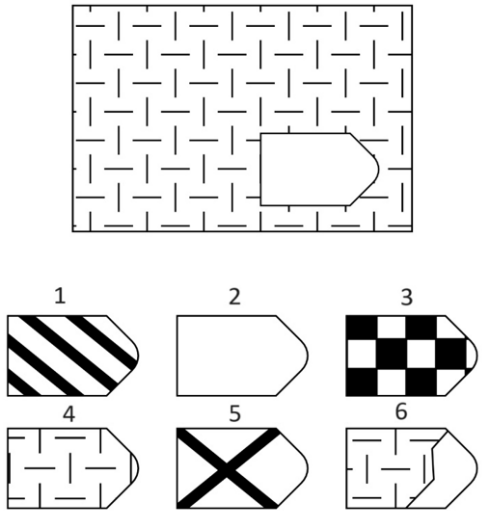


Figure OA.1.4: Sample screen in the analytical categorization task

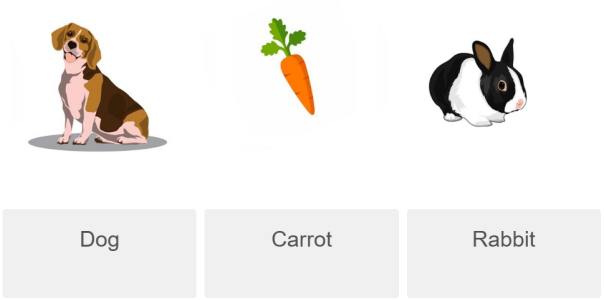


Figure OA.1.5: Sample screen in the analogical reasoning task

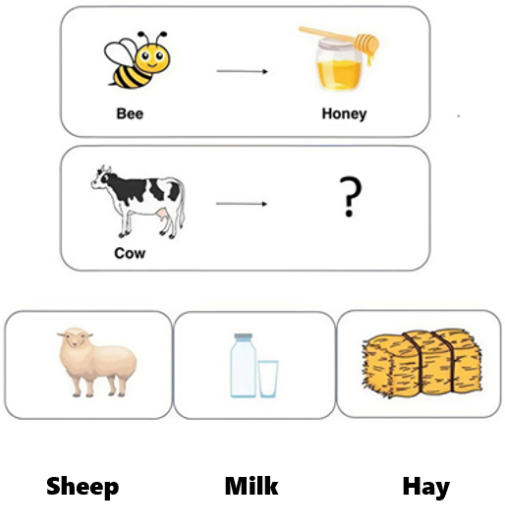
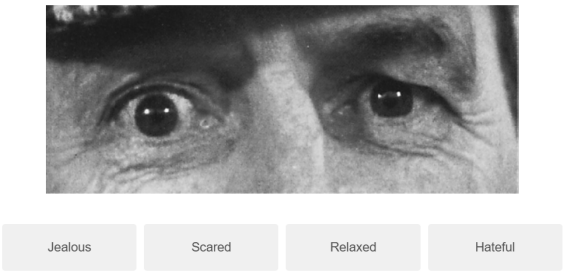


Figure OA.1.6: Sample screen in the cognitive empathy test



OA.1.2 Risk Game

In the risk-preference task, each student receives an endowment of 5 tokens and chooses how many tokens, an integer between 0 and 5, to allocate to a risky option. Tokens allocated to the risky option are subject to a lottery with two equally likely outcomes. A ball is drawn blindly at random from a bag containing exactly two balls, one yellow and one purple. If the yellow ball is drawn, the tokens allocated to the risky option are tripled; if the purple ball is drawn, the allocated tokens are lost (set to zero). Tokens not allocated to the risky option are kept with certainty regardless of the lottery outcome. Students make their allocation decision before the draw and do not know which ball will be drawn at the time of choice. Therefore, for a risky allocation of $x \in \{0, 1, 2, 3, 4, 5\}$, payoffs in tokens are:

$$\text{Payoff}(x) = \begin{cases} 2x + 5, & \text{if the yellow ball is drawn,} \\ 5 - x, & \text{if the purple ball is drawn.} \end{cases}$$

Since the two outcomes occur with equal probability, the expected payoff is:

$$E[\text{Payoff}(x)] \;=\; \tfrac{1}{2}(2x + 5) \;+\; \tfrac{1}{2}(5 - x) \;=\; 5 + \tfrac{x}{2}.$$

which is increasing with x.

Instructions:

Hello everyone! We will play a fun game with you today. By playing this game, you will have a chance to earn gift points. Each point corresponds to a gift of equal value from our gift bag [show the gift bag to students]. The decisions you make during the game will determine the number of gift points you earn. The more points you earn, the more and nicer gifts you get!

In this game, each of you receives 5 tokens. These tokens are yours for sure. You can also place some of your tokens in our chance box [draw a chance box on the board]. Let me explain what happens to the tokens you put into the chance box. I will draw one ball from this black bag without looking. There are only two balls in the bag, the same size but different colors, one yellow and one purple [show the two balls to students clearly and place them back into the bag]. If the yellow ball is drawn, the tokens you put into the chance box are tripled. If the purple ball is drawn, you lose the tokens you put into the chance box. Tokens kept outside the chance box remain yours, regardless of which ball is drawn.

Let us try to understand the game with some examples:

1. If you put 1 token into the chance box:
   - If the yellow ball is drawn: $1 \times 3 = 3$ in the chance box and 4 tokens stay safe, so you earn $3 + 4 = 7$ tokens.
   - If the purple ball is drawn: you lose the 1 token in the chance box, and 4 tokens stay safe, so you earn 4 tokens.
2. If you put 2 tokens into the chance box:
   - If the yellow ball is drawn: $2 \times 3 = 6$ in the chance box and 3 tokens stay safe, so

you earn $6 + 3 = 9$ tokens.

- If the purple ball is drawn: you lose the 2 tokens in the chance box, and 3 tokens stay safe, so you earn 3 tokens.

3. If you put 3 tokens into the chance box:

   - If the yellow ball is drawn: $3 \times 3 = 9$ in the chance box and 2 tokens stay safe, so you earn $9 + 2 = 11$ tokens.
   - If the purple ball is drawn: you lose the 3 tokens in the chance box, and 2 tokens stay safe, so you earn 2 tokens.

4. If you put 4 tokens into the chance box:

   - If the yellow ball is drawn: $4 \times 3 = 12$ in the chance box and 1 token stays safe, so you earn $12 + 1 = 13$ tokens.
   - If the purple ball is drawn: you lose the 4 tokens in the chance box and 1 token stays safe, so you earn 1 token.

5. If you put all 5 tokens into the chance box:

   - If the yellow ball is drawn: $5 \times 3 = 15$, so you earn 15 tokens.
   - If the purple ball is drawn: you lose all 5 tokens, so you earn 0 tokens.

6. If you do not put any tokens into the chance box, it does not matter whether a yellow or a purple ball is drawn; you get 5 tokens at the end of the game.

OA.1.3 Critical Thinking Game

The critical thinking game presents students with a short paragraph describing a lost item in a hypothetical school setting, together with six possible locations, only one of which is correct. Students earn 10 tokens if they identify the correct location, and 0 otherwise. Each student also receives an endowment of 6 tokens that can be used to purchase clues. Each clue is presented as a short paragraph and provides one piece of information that rules out exactly one of the six locations. Each clue costs 1 endowment token, and students may purchase up to four clues. Purchasing four clues reduces the set of plausible locations from six to two, implying a 50% chance of selecting the correct location if the student guesses between the remaining options.

After reading the initial mysterious story paragraph, students decide whether to purchase a clue. If a student chooses not to purchase any clues, they proceed directly to select one of the six locations as their final response. After making a choice, the student is asked why they decided not to purchase a clue and selects from predefined response options: "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 purchases a clue, one endowment token is deducted, and the student is shown the first clue paragraph. The clue implicitly indicates one location where the lost item cannot be. The student is then asked to identify which of the six locations is ruled out by the clue. After eliminating one option, the student is asked to state what the clue implies, capturing their ability to justify the elimination. Students respond by choosing among three predefined statements describing the clue's implication. These response options are intentionally designed to be confusable to assess whether students can accurately articulate the reasoning underlying their elimination decision. The student then decides whether to purchase another clue or stop and select a location. This sequence repeats for up to four clues, depending on the student's decisions.

Besides the clue paragraphs, the initial storyline provides no information favoring one location over another. Accordingly, conditional on the information revealed up to a given stage of the task, all remaining locations are equally likely to be the correct answer. At each stage, only one location yields the prize of 10 tokens, while the other locations yield zero prize tokens. Let $x \in \{0, 1, 2, 3, 4\}$ denote the number of clues purchased by student i. Purchasing clues reduces the student's remaining endowment from 6 to $6 - x$. Thus, the student's total payoff in tokens equals

$$\text{Payoff}_i(x) = (6 - x) + 10 \cdot \mathbb{1}\{\text{correct location is chosen}\}.$$

Equivalently, payoffs are $10 + (6 - x)$ if the student chooses the correct location, and $6 - x$ otherwise.

If each purchased clue is used correctly to eliminate exactly one incorrect location, then after purchasing x clues, the number of remaining candidate locations is $6 - x$. Under the assumption that the student selects randomly among the remaining locations when making

the final choice, the probability of choosing the correct location is

$$\Pr(\text{correct} \mid x) = \frac{1}{6-x}.$$

Under this benchmark, the expected payoff as a function of clue purchases is

$$\mathbb{E}[\text{Payoff}_i(x)] = (6-x) + 10 \cdot \frac{1}{6-x}.$$

where x denotes the number of clues purchased by student i. This formula of expected payoff highlights the trade-off inherent in the task: purchasing additional clues reduces the remaining token endowment but increases the probability of selecting the correct location. The expected payoff is not monotone in x. In particular,

$$\begin{aligned}
\mathbb{E}[\text{Payoff}_i(0)] =& 6 + \frac{10}{6} \approx 7.67, \\
\mathbb{E}[\text{Payoff}_i(1)] =& 5 + \frac{10}{5} = 7, \\
\mathbb{E}[\text{Payoff}_i(2)] =& 4 + \frac{10}{4} = 6.5, \\
\mathbb{E}[\text{Payoff}_i(3)] =& 3 + \frac{10}{3} \approx 6.33, \\
\mathbb{E}[\text{Payoff}_i(4)] =& 2 + \frac{10}{2} = 7.
\end{aligned}$$

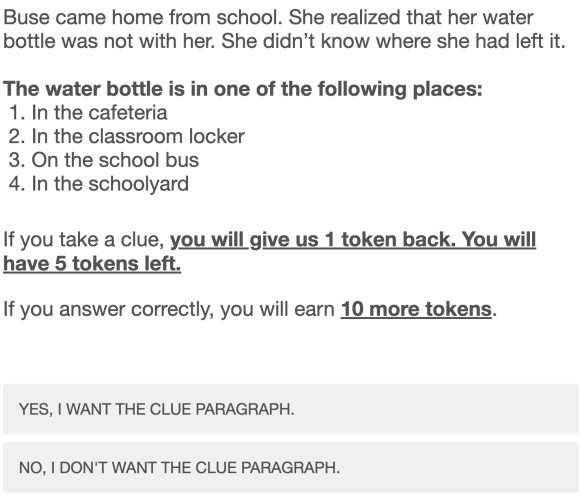
Therefore, under this benchmark, purchasing additional clues is not payoff-maximizing until the fourth clue; the expected payoff increases only when moving from three to four clues.

Given the relatively complex structure of the task and its payoff scheme, enumerators first conduct a practice round with the entire classroom before the individual task begins. In this practice round, students solve a sample scenario with a short mystery story, four possible locations, and the option to purchase up to two clues. The practice round is used to ensure that all students understand the rules and the consequences of purchasing clues. The detailed instructions for administering this practice round are provided below.

Instructions:

Now, we will play another game. As with the previous one, playing this game gives you a chance to earn gift points. Each point corresponds to a gift of equal value from our gift bag [show the gift bag to students]. The decisions you make during the game will determine the number of gift points you earn. The more points you earn, the more and nicer gifts you get!

Figure OA.1.7: The sample screen in the practice game presenting the story and the lost item



In this game, you will solve a mystery about a lost item, like detectives trying to find out what happened. At the start, each of you receives 6 tokens. These tokens belong to you, but you can choose to use some of them to help you solve the mystery. If you solve the case and find the correct location of the lost item, you earn 10 extra tokens in addition to the tokens you still have. If you do not find the correct location, you do not earn the 10 extra tokens and keep only the tokens you have left.

To make sure everyone understands how the game works, we will first solve one mystery together as a class. This is just a practice round, so we will not win or lose tokens yet. After we finish the practice together, each of you will play the real game on your own. Now, let us read the story we are given and find out what item is lost in our story.

After reading the story, we see the possible places where the water bottle could be. In this practice round, there are four options. Right now, we do not know where the bottle is. It could be in any of the four places, so choosing the correct one is not easy when we have four options.

Now, you have a choice. If you want, you can buy a clue to help you solve the mystery. A clue is a short paragraph. The clue does not tell you exactly where the bottle is, but it helps you by showing you one place where the bottle cannot be. So, if you buy one clue, you can eliminate one option. Then you will have three options left, and it becomes easier to choose the correct one. But the clue is not free. To buy one clue, you must give back one

of your six tokens.

After the first clue, you can decide again. If you want, you can buy a second clue. If you buy a second clue, you give back one more token. This second clue helps you eliminate one more option. Then only two options remain, and you are closer to the correct answer.

Let us try to understand how we can play with some examples:

- Let us assume one of our classmates does not buy any clue. That student keeps all 6 tokens and must choose among 4 options. If the student finds the correct location, they earn 10 extra tokens and finish with 16 tokens. If the student does not find it, they finish with 6 tokens.

- Let us consider another classmate who wants to buy only one clue. That student spends 1 token, so 5 tokens remain. One option is eliminated, and the student chooses among 3 options. If the student finds the correct location, they finish with 15 tokens; if not, they finish with 5 tokens.

- A third classmate buys two clues. That student spends 2 tokens, so 4 tokens remain. Two options are eliminated, and the student chooses among 2 options. If the student finds the correct location, they finish with 14 tokens; if not, they finish with 4 tokens.

These are only examples to understand the rules. In the practice round, we do not actually win or lose tokens.

Now, let us try the first clue together (see Figure OA.1.8). Based on this clue, where can the bottle not be? It cannot be on the bus, because the driver said he did not see it. So we eliminate the bus option. Now one option is gone, and three options remain.

Next, the game asks: "What was the clue in the paragraph?" In other words, what information helped us eliminate the bus? The important point is that Buse asked the driver, and he said he did not see any bottle on the bus. That is why we can say the bottle is not on the bus (see Figure OA.1.9).

Now we are ready to move to the real game. In the real game, you will again start with

Figure OA.1.8: The sample screen in the practice game presenting the first clue

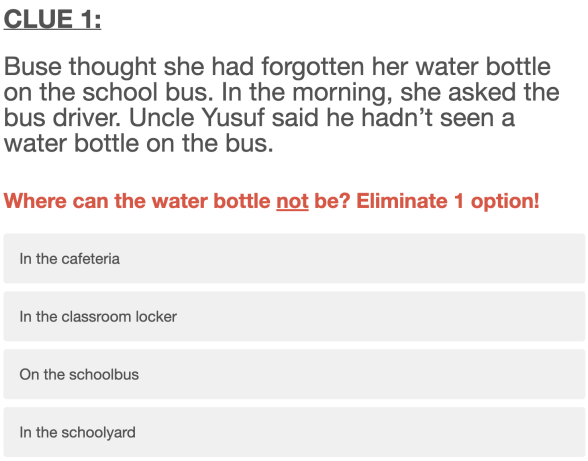
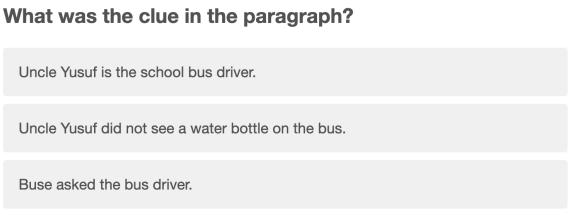


Figure OA.1.9: The sample screen in the practice game presenting the justification of the previous elimination



6 tokens, but this time there will be six possible locations. You will try to find the lost item among these six options. Also, this time you can buy up to 4 clues.

OA.2 Program Implementation

Figure OA.2.1: Book covers



Program Materials and some examples

Figure OA.2.2: Teacher Training Sessions



Figure OA.2.3: In class Implementation

