



## OPEN ACCESS

### EDITED BY

Yaranay López-Angulo,  
University of Concepcion, Chile

### REVIEWED BY

Ahmad Lutfi Fauzi,  
Indonesia University of  
Education, Indonesia  
Fatia Fatimah,  
Indonesia Open University, Indonesia

### \*CORRESPONDENCE

Roberto Araya  
✉ roberto.araya.schulz@gmail.com

RECEIVED 25 November 2025

REVISED 14 February 2026

ACCEPTED 23 February 2026

PUBLISHED 27 March 2026

### CITATION

Danoebroto SW, Wahyudi W, Ulloa O,  
Rizky R, Syaifuddin A and Araya R (2026)  
Unleashing computational thinking: a  
novel approach to reforming Indonesian  
primary math instruction.  
*Front. Educ.* 11:1754373.  
doi: 10.3389/feduc.2026.1754373

### COPYRIGHT

© 2026 Danoebroto, Wahyudi, Ulloa,  
Rizky, Syaifuddin and Araya. This is an  
open-access article distributed under the  
terms of the [Creative Commons  
Attribution License \(CC BY\)](#). The use,  
distribution or reproduction in other  
forums is permitted, provided the  
original author(s) and the copyright  
owner(s) are credited and that the  
original publication in this journal is  
cited, in accordance with accepted  
academic practice. No use, distribution  
or reproduction is permitted which does  
not comply with these terms.

# Unleashing computational thinking: a novel approach to reforming Indonesian primary math instruction

Sri Wulandari Danoebroto<sup>1</sup>, Wahyudi Wahyudi<sup>1</sup>, Obed Ulloa<sup>2</sup>,  
Rizky Rizky<sup>3</sup>, Akhmad Syaifuddin<sup>4</sup> and Roberto Araya<sup>2\*</sup>

<sup>1</sup>Southeast Asian Ministers of Education Organization (SEAMEO) Regional Centre for QITEP in Mathematics, Yogyakarta, Indonesia, <sup>2</sup>Centro de Investigación Avanzada en Educación, Instituto de Educación, Universidad de Chile, Santiago, Chile, <sup>3</sup>University of AMIKOM, Yogyakarta, Indonesia, <sup>4</sup>Universitas Sebelas Maret, Surakarta, Indonesia

There are two growing demands in primary mathematics education: to deepen students' mathematical thinking, and to integrate Computational Thinking (CT)—the conceptual and algorithmic foundation underpinning artificial intelligence. However, current pedagogical strategies have proven insufficient in effectively bridging both demands. To address this persistent gap, we propose an innovative instructional syntax centered on structured, metacognitive reflection across three levels: individual, peer-to-peer, and collective. The syntax organizes argumentation, coloring tasks, creative problem posing, peer interaction, and clustered collective reflection. Students perform these actions in written form, creating a traceable record that facilitates systematic analysis and enables comparisons among peers and across the whole class. This syntax facilitates dialogical learning, ensuring all students actively participate. We piloted this instructional syntax in 17 classes with 215 primary students and examined implementation fidelity using LLM-supported transcript analysis. The implementation was effective, demonstrating the feasibility of integrating this syntax into regular classroom practice. To the best of our knowledge, this analysis represents the first application of Large Language Models (LLMs) and clustering techniques to understand the implementation of such a dialogic and reflective instructional approach. The LLM-based analysis of the lesson transcriptions, focused on the instructional syntax, provides clear guidance for continuous improvement and scaling the integration of both demands.

### KEYWORDS

computational thinking, dialogic intelligence, inner voice, LLMs based analytics, mathematical discourse, metacognition, primary mathematics education

## 1 Introduction

The rapid advancement of technology and the growing prevalence of Artificial Intelligence (AI) in everyday life create a unique and complex landscape for preparing children for the future. According to the recent Nobel Prize in Economics, the AI revolution is expected to significantly increase aggregate productivity growth, which, depending on the calculation method, is estimated to be between 0.68 and 1.3 percentage points per year over the next decade (Aghion and Bunel, 2024). Thus, traditional educational models, which often prioritize rote memorization and procedural knowledge, are becoming less relevant

as AI systems can perform these tasks faster and more accurately. This shift requires a focus on skills that complement, rather than compete with, AI, specifically those enabling critical evaluation and responsible deployment of these powerful tools.

One of the most crucial of these skills is Computational Thinking (CT). CT is not about learning to code, but rather about approaching problems in a way that can be understood and solved by a computer. It involves breaking down complex problems into smaller, more manageable parts (decomposition), recognizing patterns in data (pattern recognition), developing step-by-step instructions to solve the problem (algorithms), and generalizing solutions for a wider range of similar problems (abstraction) (Wing, 2006). CT is the thought process required to do computational models of the world (Denning, 2017; Tedre and Denning, 2016).

However, the successful and effective integration of AI into societal structures represents a profound cognitive and educational challenge. The increasing use of AI and LLMs presents citizens with new cognitive challenges. The critical hurdle is cultivating the analytical judgment needed to assess the reliability and validity of AI-generated content. Using AI and LLMs without understanding their internal mechanisms and without developing computational thinking skills poses several risks, as LLMs exhibit remarkable generative capabilities but remain vulnerable to complex failures (Kazlaris et al., 2025; Kargupta et al., 2025). These models are susceptible to unpredictable hallucinations—outputs that are fluent yet inaccurate, ungrounded, or inconsistent with source material. Furthermore, due to their training data and pattern-matching architecture, LLMs can exhibit functional analogs of human fallibility, including cognitive biases and illusions (Araya, 2025b), as well as statistical aversions or “phobias” (Araya, 2025a). This inherent susceptibility to error demands that users develop sophisticated metacognitive skills to calibrate their trust in AI suggestions and outputs. Understanding these vulnerabilities is critical for responsible deployment. By nurturing CT from an early age, we can empower children to become creators and innovators in a world increasingly shaped by technology, capable of harnessing AI’s power while mitigating its inherent risks. This skill set prepares them not just for careers in STEM but also for a future in which technology is a fundamental tool across every field.

Several studies and data from international assessments and competitions highlight weaknesses in the instruction of mathematical and computational thinking in Indonesian primary schools. Several studies support the critical need for pedagogical innovation. We review several of them.

First, let’s consider student Proficiency studies in mathematics. The Programme for International Student Assessment (PISA) 2018 report revealed that Indonesian students’ mathematical literacy was significantly low, with an average score of 379, well below the OECD average of 489. Indonesia’s PISA 2022 results show declines across the three domains—mathematics, reading, and science—relative to earlier assessments. In 2022, mean scores were among the lowest Indonesia has recorded: mathematics (~366), reading (~359), and science (~383). These results are similar to those in the early PISA years (e.g., 2003 for math/reading) and indicate a decline from achievements in the mid-2010s (OECD PISA 2022 Results (Volume I and II)—Country Notes: Indonesia) (OECD, 2023). Although the gap between high and low achievers in mathematics

has narrowed recently, it is mostly because top performers’ scores have dropped, while low performers have shown smaller changes. In reading and science, however, the distribution didn’t change much: disadvantage persists. Socioeconomically advantaged students outperform disadvantaged students in mathematics by ~34 score points in Indonesia, though this is far less than the OECD average gap (~93 points). The influence of socio-economic status on mathematics performance in Indonesia is smaller than the OECD average, accounting for approximately 6% of the variability.

Second, we review CT. Between 2019 and 2025, research on integrating CT into Indonesian mathematics classrooms has increased, often employing task-based approaches such as project-based learning (PBL) or Realistic Mathematics Education (RME). Many studies still remain qualitative and descriptive; only a few use experimental or intervention designs (Suarsana et al., 2024); programming tools (e.g., block-based or tangible devices) are infrequently used; and the literature shows that many CT implementations are carried out without actual coding tools (Suarsana et al., 2024). The national curriculum formally includes CT as a “minimum competency” from elementary through high school, often embedded in mathematics, science, and informatics subjects (Suarsana et al., 2024). According to UNICEF’s “Digital Learning Landscape in Indonesia” (UNICEF, 2021), a major challenge is the limited adoption of educational technology by teachers and students, as many are unaware of government digital learning platforms, and integration into regular teaching is low. Moreover, unequal digital access (rural/urban, socioeconomic divides) constrains equitable CT education (SMERU Research Institute, 2022; UNICEF, 2021). In sum, while CT is endorsed in policy and increasingly researched, the implementation in programming-based, creative CT activities remains modest and uneven in Indonesia, hampered by resource, access, and capacity gaps.

A study on the Bebras Challenge 2023, an international CT competition (Fitriyah et al., 2024), found that 87% of Indonesian participants scored below the minimum proficiency threshold of 50. This indicates that the vast majority of students lack fundamental CT skills. On the other hand, research on students’ CT skills in mathematics reveals consistent challenges across grade levels and contexts. For instance, Rosali and Suryadi (2021) found that while one eighth-grade student scored 67.48 on a CT test—classified as “high”—the study still highlighted significant gaps in core CT components such as abstraction and algorithmic thinking. Similarly, Permana et al. (2022) emphasized the difficulty that elementary students face when applying CT strategies in mathematical problem-solving, despite the use of structured interventions. Fajri and Yurniwati (2019) further demonstrated that students’ cognitive styles play a critical role in shaping their CT and mathematical thinking abilities, suggesting that instructional approaches must be tailored to individual learning profiles.

A recent adjustment (2023–2024) to the Merdeka Curriculum reintroduced informatics as a subject in many schools, emphasizing not only digital literacy but also logical reasoning, computational thinking, and the distinction between fact/opinion. It focuses on developing CT skills, fostering logical reasoning for independent problem-solving (GEM Report UNESCO, 2023). However, it faces several challenges.

What about creative thinking? According to the [OECD \(2024\)](#) PISA Results (Volume III)—Factsheets: Indonesia, Indonesia's mean score in creative thinking was 19 points out of 60, well below the OECD average of 33. Only 31% of Indonesian students reached the baseline proficiency (Level 3) in creative thinking, compared to an average of 78% across OECD countries. Only 5% of Indonesian students are top performers in creative thinking (Levels 5–6), which is significantly lower than in many peer countries. According to the OECD report's "pedagogies conducive to creativity" section, approximately 18% of schools in Indonesia offer computer programming classes or activities at least weekly. Among students, 31% attend them. Despite this, creative thinking performance remains low, and Indonesia exhibits larger gaps: many students do not reach the baseline level of creative thinking proficiency, even with exposure to activities such as programming.

Third, we review teacher Preparedness and Training. According to a UNESCO paper commissioned for the [GEM Report UNESCO \(2023\)](#), "Technology in Education in Southeast Asia," there is a deficiency in teacher training on the effective utilization of technology in the classroom. Without these foundational elements, the transformative potential of this curriculum cannot be fully realized across the nation's diverse educational landscape. The main challenge of the Kurikulum Merdeka has been relying on teachers' competencies, requiring them to shift their paradigms and facilitate students' active learning ([GEM Report UNESCO, 2023](#)). A survey of 100 teachers in Indonesia ([Asian Development Bank, 2022](#)) revealed that while 70% had access to technology, only 45% felt confident using it as an educational tool. This indicates a major gap between technological access and the pedagogical skills needed to leverage it effectively. The same survey found that 85% of teachers acknowledged they had not received specific training on how to use technology in education. These data point to a systemic failure in providing educators with the necessary skills to integrate CT-based instruction.

A World Bank study ([Dini et al., 2024](#)) on teaching practices in Indonesian primary schools found that teachers scored poorly on key instructional elements. For example, they scored 2.4 out of 5 for encouraging students to "think critically," a core component of CT, and 1.8 out of 5 for providing effective feedback. This data shows that current teaching methods are not effectively fostering CT skills.

On the other hand, to achieve the goal of integrating computational and mathematical thinking, any novel approach must consider the dynamics of cognitive processes during learning. Additionally, it should strive to incorporate recent AI developments that can support it. One fundamental fact is that human cognition is a complex interplay of two different cognitive systems ([Kahneman, 2011](#); [Stanovich et al., 2016](#)), each contributing to how we process information and solve problems. The dual-process theory posits two modes of thought: System 1 and System 2. System 1 is fast, automatic, and intuitive, operating without conscious effort. It's the system we use for tasks like recognizing a face or performing simple arithmetic, such as  $2 + 2$ . Conversely, System 2 is slow, deliberate, and effortful, requiring conscious attention to solve complex problems, such as a difficult math equation or a logical puzzle. This system is what we engage when we are actively monitoring our thoughts and reasoning through a problem step

by step. The distinction between these two systems provides a powerful framework for understanding not only human decision-making but also the behavior of artificial intelligence. Key activities, such as argumentative writing, problem-solving, and problem-posing, are not merely tasks; they require reflection on our internal cognitive processes at work. When we write an argument, we actively structure our thoughts, identify premises, and formulate conclusions. Similarly, solving a problem requires us to navigate a series of steps, and posing a problem involves a deep understanding of a concept to the point where we can formulate new questions about it. These activities highlight the dynamic and multifaceted nature of our mental operations, which are composed of two distinct systems ([Kahneman, 2011](#); [Stanovich and Toplak, 2023](#)).

Interestingly, artificial intelligence (AI) exhibits behaviors that resemble those of cognitive systems ([Fabiano et al., 2025](#); [Joseph, 2025](#); [Bengio, 2004](#); [Guo et al., 2025](#); [Street et al., 2024](#)). While most AI models today operate on a more structured, algorithmic basis, some advanced models are beginning to show characteristics of System 1. They can process vast amounts of data and make rapid, intuitive-like decisions, sometimes without a transparent step-by-step reasoning process. This is particularly evident in models that have a degree of autonomy or can generate novel content. For instance, a generative AI that creates a piece of art or writes a story may be applying a System 1-like process, relying on learned patterns and associations to produce a result that feels fluid and automatic, rather than a rigid, System 2-like calculation.

A critical aspect of human cognition is the phenomenon of our inner voice or self-talk ([Kross, 2021](#)). This internal monolog is more than just a running commentary; it is a crucial mechanism for monitoring and controlling our thought processes ([Moser et al., 2017](#); [Roby and Kidd, 2008](#)). It allows us to review our ideas, correct errors, and stay on track toward a solution. This self-monitoring function is a manifestation of System 2 thinking, where we consciously reflect on our cognitive operations. By "talking to ourselves," we engage in a feedback loop that enhances our ability to reason and solve complex problems. This inner dialogue is a testament to the metacognitive abilities that are fundamental to advanced human thought. Thus, to foster argumentation and posing problems, we propose an "embodied argumentation" strategy ([Araya et al., 2025a](#)). This approach utilizes a hand puppet to facilitate the internalization process described by [Vygotsky \(1978\)](#) and [O'Connor \(2020\)](#). The process transforms external social dialogues into internal inner-speech dialogues. This essential process—moving from social to private to inner speech—transforms an interpersonal activity into a deep intrapersonal cognitive skill.

Moreover, the effectiveness of inner speech and embodied argumentation depends heavily on how teachers orchestrate classroom dialogue. Monitoring and fostering dialogic interactions is, therefore, critical, since such exchanges provide the raw material that students gradually internalize into private and inner speech. Dialogic pedagogy emphasizes coherence, authorship, and responsible talk as pathways for deeper reasoning and collective meaning-making ([Lehesvuori et al., 2013, 2023](#); [Alexander, 2020](#); [Lehesvuori and Ametller, 2021](#)). Interaction and dialogic practices, with the teacher or others, foster the development of argumentation skills ([Kuhn, 2015](#); [Iordanou and Rapanta,](#)

2021). Thus, situating argumentation within dialogic structures, teachers create opportunities for students to practice reflection and perspective-taking, skills that eventually become part of their inner cognitive toolkit for problem-solving, problem-posing, and creative inquiry.

Metacognition is a reflection and conversation with one’s inner voice about how one approaches and solves a problem. Surprisingly, this is also critically important in AI. New Large Language Models (LLM) are incorporating it, given their greater effectiveness in reasoning (Bengio, 2004; Didolkar et al., 2025; Araya, 2025a; Guo et al., 2025). Inner speech saves memory for reasoning. In other words, it requires less cognitive load. During the reasoning process, LLMs often reroute themselves through the same intermediate steps between problems, which increases token usage and latency. This saturation of the context window reduces exploration capacity. Therefore, the inner speech strategy proposes a metacognitive analysis of the model, revisiting traces of previous chains of reasoning. Thus, the model improves its own future reasoning by leveraging behaviors from its previous problem-solving attempts. LLMs learn to evaluate their own ongoing reasoning by writing and reflecting on the statements they’ve already generated, and to explore alternative approaches in their responses. For example, the model learned to insert phrases into their reasoning, such as: “Wait. That’s a revelatory moment I can point to here” (Ippolito and Zhang, 2025), or “Wait, but that seems contradictory. Maybe the better way is to explain through truth tables” (Araya, 2025b).

Finally, pedagogical tools such as the use of color play a significant role in making complex, abstract concepts more concrete and accessible. This is particularly powerful in abstract fields such as mathematics. Color helps students understand abstract ideas by providing a visual anchor. For example, using different colors to represent various geometric shapes or to highlight different parts of an equation can make the information easier for our System 1 to process, thereby freeing up System 2 for higher-level reasoning. This illustrates how external aids, like visual cues, can be strategically employed to bridge the gap between abstract concepts and our cognitive systems, making the learning process more intuitive and effective (Araya and Isoda, 2023; Araya, 2021a, 2025c; Somsaman et al., 2024). These types of multimodal teaching strategies, which utilize drawings and colors, such as the kawung batik motif (Danoebroto et al., 2024), have shown promise in developing students’ motivation and mathematical thinking. They foster the recognition of patterns, abstract and analogical thinking, functional reasoning, and symbolic representation, and can do so by cultivating diverse dimensions of mathematical thought in culturally meaningful ways.

The coloring activities in the A/B worksheets promote diverse CT components, spanning numerical, fractional, spatial, logical, recursive, inferential, probabilistic, and statistical reasoning, as well as mathematical and computational modeling, as detailed in Table 1. These tasks bridge fundamental CT concepts, ranging from Turing machines with varying instruction sets to Von Neumann cellular automata (Agüera y Arcas, 2025a,b; Changsri et al., 2025). By modeling biological entities—such as bacteria or worms as “Pac-mans” navigating boards via gradient descent methods (Araya, 2021b, 2023)—the activities provide a rigorous but playful framework for understanding life phenomena and organic behavior. This coloring approach attempts to enable primary

TABLE 1 CT components included in the A/B coloring worksheets and illustrative examples.

Components	Example
Fractional thinking	Paint red two-thirds of balls of each box
Recursive thinking	Paint blue all boxes that contain boxes
Spatial thinking	Paint yellow all cats under a chair
Logical thinking	Paint yellow the smallest chick belonging to the largest hen
Statistical thinking	Paint black all cats above average size
Probabilistic thinking	Given the cues, paint red the book most probably contains the card
Inferential thinking	Find the appropriate sequence of clues and by following it infer the object where the card is hidden, and paint that object red.
Causal thinking	In each box, paint each ball the color of the ball in the position directly above but in the adjacent left box.
Planning	Paint where the ball will land
Mathematical modeling	Paint in red the path of a worm that starts at the top left of the board and moves each time to the adjacent cell with more nutrients, indicated by the number inside the cell.
Computational modeling	Paint in red the path of a worm that starts at the top left of the board and moves each time to the adjacent cell with more nutrients, indicated by the number inside the cell.
Pattern recognition	Paint the balls in each box with the opposite pattern of the nearest box
Written argumentation	Explain with your own words, using a story about pets.
Problem posing	Pose a similar problem, but not exactly the same as the previous one
Metacognition	Using the words typical of your model of your peer, explain how your peer understands your sentence
Prompting	For each object, select two different colors from red, yellow, and blue, and paint half of the object one color and the other half the other color.
Algorithmic thinking	Select a number between 2, and 4, in each box choose a color and paint at least that amount of that color.

students to internalize fundamental computational abstractions through embodied, visual execution.

While previous studies in Indonesia—such as Rosali and Suryadi (2021), who examined students’ CT skills in number pattern lessons during the COVID-19 pandemic, and Permana et al. (2022), who explored the integration of CT in elementary mathematics instruction—have identified persistent challenges in algorithmic reasoning, abstraction, and decomposition, comparative research across different educational systems provides valuable contrast. For instance, a cross-cultural investigation by

Prahmana et al. (2024) revealed that Indonesian and Japanese students exhibited distinct heuristic strategies when solving geometry problems, reflecting variations in curricular design and instructional scaffolding. Furthermore, a systematic literature review by Fauzi et al. (2024) emphasized that successful integration of CT occurs more effectively when curricula, teacher professional development, and assessment systems are cohesively aligned with CT principles. This juxtaposition highlights two critical gaps in the Indonesian context: (a) a lack of fine-grained, process-level data on how teachers scaffold CT during classroom discourse, and (b) insufficient empirical linkage between teacher–student interaction patterns and quantifiable CT constructs. These gaps directly inform the present study’s methodology, which employs instructional syntax analysis and artificial intelligence (AI)-based discourse mapping to uncover latent CT structures embedded in teacher–student communication. By analyzing classroom talk as computational data, this approach transcends conventional pre- and post-test measures, offering a dynamic, process-oriented understanding of CT development within authentic learning environments.

A key methodological innovation in our work is the use of AI, and particularly LLMs, to analyze classroom practices (Urrutia and Araya, 2024; Tapia-Mandiola and Araya, 2024). Recent studies, such as those by Lehesvuori et al. (2025), have demonstrated that teacher questions can be automatically analyzed from classroom transcripts, highlighting the potential of technology-aided discourse research. Similarly, Xu et al. (2024) evaluated ChatGPT and GPT-4 for coding classroom discourse in online mathematics instruction, demonstrating the reliability of LLMs in handling complex interaction data. There is also evidence that AI can help estimate the effect of teacher discourse on student learning outcomes. For example, Schlotterbeck et al. (2020) found discourse patterns linked to measurable learning gains. Furthermore, AI can reconstruct teachers’ conceptual networks directly from transcripts (Caballero et al., 2017). Together, these advances demonstrate that AI-based analysis provides scalable, replicable, and in-depth insights into teaching and learning.

In this paper, we address two research questions:

**RQ 1.** How can an instructional syntax of CT integrated with mathematics teaching be applied in Indonesian elementary schools?

**RQ 2.** What specific patterns, identified through AI analysis of lesson plans and classroom videos, can be used to pinpoint activities to increase the effectiveness of students’ CT learning?

## 2 Methods

The methodology centers on the systematic application and monitoring of an innovative instructional syntax designed to foster structured, metacognitive reflection across individual, peer-to-peer, and collective levels. This framework organizes pedagogical actions—including argumentation, coloring tasks, problem posing, and clustered reflection—into traceable written records, enabling comparative analysis of student participation and dialogical learning. To address our core research questions, the study piloted the instructional syntax in 17 primary classrooms. The analysis

assesses the degree of implementation fidelity and the practical feasibility of integrating this instructional syntax into authentic primary educational settings, using LLM-supported transcript analysis for rigorous evaluation.

The methodology has three phases. First, teachers are trained in the A/B coloring books (Araya, 2024) and the novel instructional syntax involved. Second, the pilot phase and data collection of lesson plans, video recordings of the sessions, and student work. Third, the analysis of the collected data. In this paper, we focus on the alignment between the proposed syntax and lesson plans based on actual classroom teaching.

### 2.1 Phase 1: teacher training

During a three-day in-person training session on 16–18 October 2024, 30 elementary school teachers from three different districts participated in a workshop led by two of the paper’s authors. It was held at the SEAMEO Regional Center for Quality Improvement of Teachers and Education Personnel (QITEP) in Mathematics. The workshop, titled “Implementation of Unplugged Computational Thinking Coloring Book in Bahasa Version for Primary School Teachers,” was designed to improve teacher competence in line with the mission of the SEAMEO Regional Center for Quality Improvement of Teachers and Education Personnel (QITEP) in Mathematics. Participants received hands-on training in A/B coloring-book activities to learn core CT concepts and how to teach them effectively. A/B coloring activities offer a new level of active and open-ended participation. Unlike traditional textbooks, A/B textbooks encourage active writing, explanation, coloring, dialogue, and problem posing (Araya, 2025c). Based on the concept of adversarial collaboration (Kahneman, 2011), students tackle the problem on page A and then innovate by posing a problem on page B to their peers (Figure 1), which encourages precise argumentation, in-depth interactions, and intensive social learning and development of prosocial skills, attitudes, and trust in peers (Araya and González, 2025). This strategy fosters creative inquiry and critical questioning, which is subsequently reviewed by the entire class. A/B coloring textbooks’ aim is to prepare students to understand and interact with Artificial Agents.

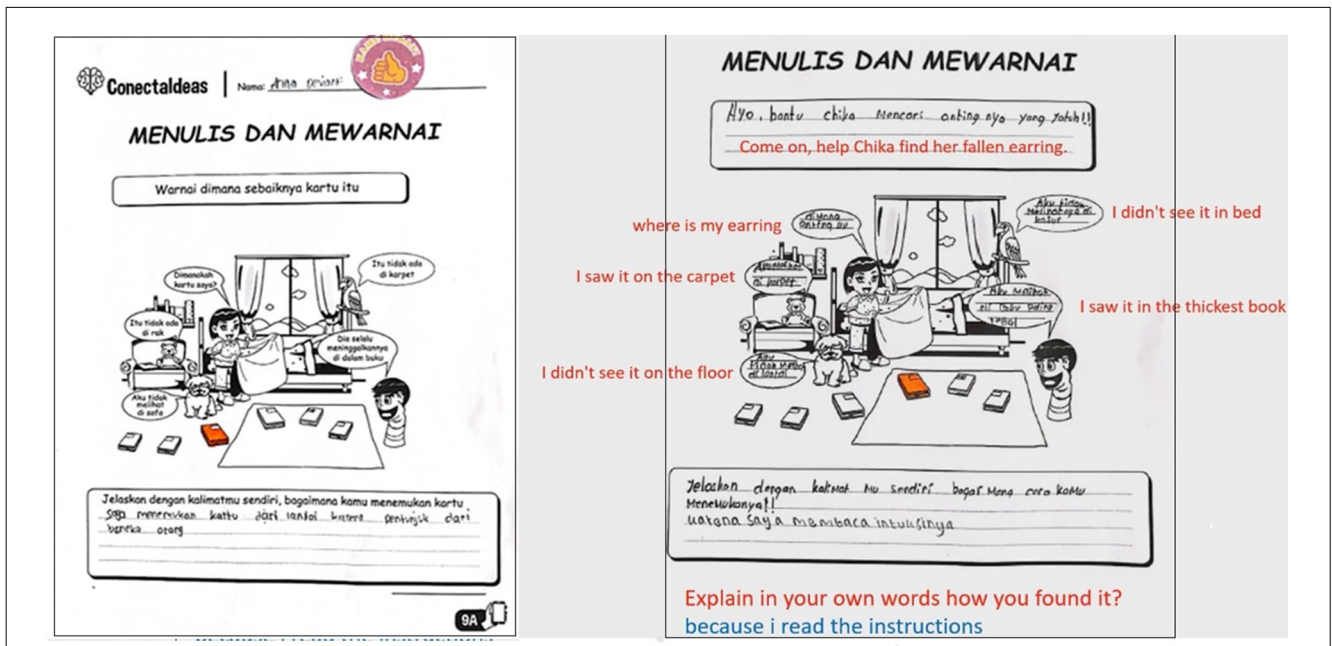
Teachers sent a total of 385 A/B coloring worksheets from the three A/B coloring books. By grade level, the number of worksheets is in Table 2.

The worksheets included activities on numerical thinking with whole numbers, fractional thinking, logical thinking with propositions and logical quantifiers, recursive thinking, and statistical and probabilistic thinking. Each teacher colored the A sheets according to the instructions and recorded their method in writing (Figure 2).

They also formulated and wrote on page B a coloring problem similar to the one on the corresponding page A, which was then completed by a nearby teacher (Figure 3).

Participants also learned about using hand puppets as a strategy to stimulate and develop metacognition by externalizing internal dialogue (Figure 4).

The A/B Coloring Books WorkSheets define the instructional syntax. This is the WS-Syntax. It is the core of the teaching strategy.



a

b

FIGURE 1 (a) page A, where the student colors following the printed instructions. (b) page B, where the student poses instructions, and then a peer will color and explain. English translations have been added in red letters to facilitate understanding by readers of the paper.

TABLE 2 Number of worksheets collected from grade levels.

Grade	1 <sup>st</sup> Grade	2 <sup>nd</sup> Grade	3 <sup>rd</sup> Grade	4 <sup>th</sup> Grade	5 <sup>th</sup> Grade	6 <sup>th</sup> Grade	Total
Page A	41	21	92	16	5	40	215
Page B	29	15	91	20	2	13	170
Total	70	36	183	36	7	53	385



FIGURE 2 Teachers coloring the worksheets during the workshop.



FIGURE 3  
Participant explaining how they solved the problem posed by her peer.



FIGURE 4  
Participant explaining to her hand-puppet.

Basically, the A/B coloring book worksheets provide a precise structure for the session. It is a novel instructional syntax derived from the worksheet's format. Pages come in pairs: A and then B.

Each activity consists of two complementary parts, each one of one page long. Page A contains an activity with images, painting and drawing activities, and spaces for student A to write his argument.

Page B contains the same pictures, but student A uses them to pose a problem to student B, who responds by writing, drawing, and coloring.

Page A introduces core AI concepts, such as the deepest descent algorithm, as well as concepts related to logical, statistical, probabilistic, spatial, visual, algorithmic, and other forms of reasoning. It is colored and answered exclusively by student A featuring handwritten text in which she explains her reasoning and describes models and conjectures. Page B contains a problem similar to the one on page A but created and written by student A. On this page, Student A poses a problem, and Student B responds by writing explanations and coloring. Thus, once completed, page B contains a record of the interaction with student B. The interaction on page B fosters the development of prosocial attitudes and skills, including collaboration, understanding others, appreciation of a partner's ideas, and building trust (Araya and González, 2025). LLM models can help the teacher analyze the handwritten responses and images colored by both students. They should assist the teacher in conducting formative assessment and providing feedback to the students. They also provide valuable feedback to the teacher for future lessons.

The novel instructional syntax for a lesson is the WS instructional syntax provided in the *Colorea Ideas A/B* books, which consists of a sequence of 7 stages (Figure 5):

- **WS1**—Prelude;
- **WS2**—Page A: Coloring;
- **WS3**—Page A: Presentation and Discussion;
- **WS4**—Page B: Coloring;
- **WS5**—Page B: Presentation and Discussion;
- **WS6**—Reflection and Closing;
- **WS7**—Epilog.

In WS1, the teacher presents a motivation for the activity. It is suggested to do so through an anecdote or personal narrative from the teacher or some students, which connects to a real-life problem tailored to the students' interests and typical life issues.

In WS2, the teacher gives students Sheet A to respond to by coloring and writing explanations. They ensure that everyone can read the instructions. They give 5 min to color according to the instructions and to write explanations when page A asks for them.

In WS3, which lasts approximately 10 min, the teacher begins by asking who wants to show their results and describe how they did. Then they call out other students. They place the work on the board, grouping similar answers into groups or clusters. The teacher then solicits students' opinions, prompting them to reflect on the different responses represented in the different groups or clusters. The teacher asks for explanations of the main differences between the clusters. She also asks for aesthetic, linguistic, narrative, humorous, and other characteristics.

WS4: The teacher distributes Sheet B and ensures that everyone understands they must pose and write a new problem for a classmate to solve. She explains that it should not be a mere copy-and-paste of the problem on page A. He gives 10 min. Each student must write their name and the name of the classmate to whom the problem is addressed in order to solve it. The teacher emphasizes the metacognitive process, demonstrating how to converse with one's inner voice or voices, reflect on ideas from

different points of view, and consider others' opinions. A hand puppet can be used to represent the notion of metacognition and of conversing with one's inner voice, using concrete materials. The teacher demonstrates different distancing strategies using the inner voice or a puppet, such as having the puppet name the student, placing the puppet at various angles and distances, and also imagining multiple inner voices or puppets representing different characters participating in a conversation. Furthermore, each inner voice or puppet can, in turn, have its own inner voices, generating a recursive process. The teacher demonstrates how to converse with the puppet, how to establish a dialogue by analyzing different options before writing the problem, and how to do so during and after writing the posed problem. After writing and reviewing the posed problem, the students exchange worksheets. The teacher gives the corresponding pairs 5 min to solve the problem by coloring and writing explanations.

WS5: The teacher asks each pair of students to compare the instruction they wrote with their peer's response. The students have to determine if they would have performed their instruction differently on their own. Then, the teacher chooses the problem posed by a student and the pair's answer. She asks them to explain whether the answer matches what the first student thought when creating the problem, and to what extent different interpretations emerge. The teacher then asks the rest of the class to share their reflections and opinions. They repeat the process with other worksheets.

WS6: The teacher posts all the work on the board. She has everyone analyzing them. Then she invites everyone to form a line and go to the front to review the worksheets. Each student will record the one that impressed them the most, which they will then discuss. The teacher asks for feedback on the types of problems posed, creativity, aesthetics, language used, mathematical and computational thinking, and asks for reflections on peer interaction and differences in interpretation of instructions.

WS7: The teacher closes the lesson by summarizing what was done, the types of solutions and problems posed, and concluding what we have learned in the lesson.

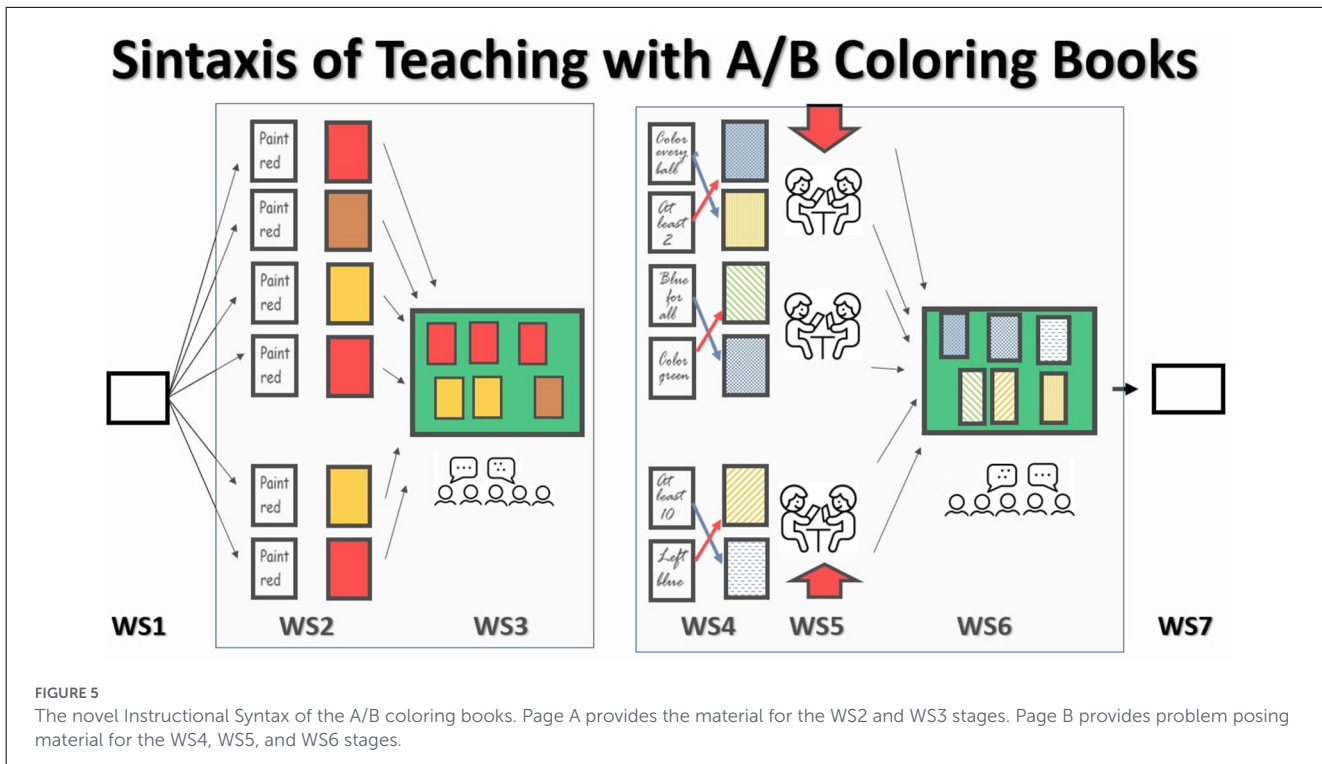
During the training, teachers practiced these activities through peer teaching and received feedback from facilitators and fellow teachers. Teachers learned together how to reflect on CT-integrated mathematics learning and trajectories of reasoning and dialogue (Schank et al., 2025).

## 2.2 Phase 2: data collection

Following the workshop, SEAMEO QITEP in Mathematics invited participating teachers to voluntarily select one of the A/B worksheet activities from the three A/B Coloring Books and pilot it following the trained strategy in a single 90-min session in their classes. 17 teachers responded and carried out the activity between November and December 2024.

Each teacher had to commit to sending three files to QITEP:

1. The Lesson Plan, a Word document that specified the chosen worksheet and the actions the teacher would perform in class.
2. The video recording of the class. It was a recording made with the teacher's camera or smartphone.



3. The scanned worksheets with the students’ work.

We collected 17 videos of these classes, taught by the 17 teachers who provided complete information. The videos belong to 17 different elementary schools, and for each one, we collected the A/B worksheets with the corresponding students’ work (Table 2). In total, this is the work of 215 students. 56% of the worksheets were problem-solving (Page A). 44% were problem-posing (Page B).

We also collected and translated the lesson plans written by each teacher, and compiled them with the videos and worksheets. From the Lesson Plans (LP), we estimated the sections and computed the time dedicated to each.

From each Lesson Plan document, which consists of identification data (date, class, allocation time, and name of the lesson), objectives, relationship with the curriculum, Learning Process, Evaluation, Source, and Tools and materials, we isolated the Learning Process component, which describes the planned structure for the lesson. We transcribed the videos using Whisper (Tigros, n.d.) with the batch tool Whisperer, which uses WhisperCpp and runs on virtually any computer. The transcribed lessons are saved as .srt files. The learning processes, along with the corresponding class transcripts, were then uploaded to Gemini. We used a two-step prompting process to analyze the data and categorize the lesson plans.

First, we prompted Gemini to align each transcript with its corresponding Learning Process of the lesson plans:

**Prompt 1:** “Can you tell me how the srt file aligns with the parts described in the Learning Process?”

After running this for all 17 pairs of files, we gave Gemini a second prompt to identify common themes:

**Prompt 2:** “Take into account all the Learning Process files provided, can you tell me which are the main stages across them?”

This process resulted in the identification of three main stages:

1. Preliminary Activities (or Introduction): This stage focuses on setting up the learning environment and preparing students for the lesson. It typically includes:
  1. Greeting students and checking their wellbeing and readiness.
  2. Praying (often led by a student).
  3. Checking student attendance.
  4. Reminders about discipline and its benefits.
  5. Apperception activities (linking to prior knowledge or setting context), sometimes involving songs or questions about daily experiences.
  6. Introducing the lesson’s objectives.
2. Core Activities: This is where the main learning and problem-solving take place, often structured using a Problem-Based Learning (PBL) model. Key steps include:
  1. Presenting a problem or scenario to students.
  2. Students engaging with the problem, observing, analyzing, and discussing.
  3. Working on worksheets (LKPD) individually or in groups.
  4. Teacher guidance and observation of student progress.
  5. Presenting and discussing results, with feedback from other groups and teacher reinforcement.
3. Closing Activities: This stage wraps up the lesson and prepares students for future learning. Common elements are:
  1. Summarizing or reviewing the material learned.

2. Reflection on the learning experience by students and teachers.
3. Providing information about the next lesson.
4. Closing with a prayer and/or greetings.

We then prompted Gemini to summarize these stages into 8 main sections of a class, giving more emphasis to the core activities. The prompt was:

**Prompt 3:** “Now consider: [8 Sections] and tell me how much time each transcript (srt files) dedicates to each of these stages.”

## 2.3 Phase 3: data analysis

### 2.3.1 Computation of alignment between WS-Syntax periods and LP periods

With  $n = 7$  sections defined by the novel instructional syntax in the WorkSheets and  $m = 8$  sections given by the Learning Process in the Lesson Plan, we compute the degree of alignment or agreement between these class segmentations. For each Instructional Syntax WS, we compute the percentage of time shared with each section of the Lesson Plan LP.

To determine the overlap percentage, we do it by computing the length of each intersection.

Given the time intervals:

$$WS_i, i \in \{1, \dots, n\}, LP_j, j \in \{1, \dots, m\}$$

Where each of them is defined by the timestamps, making the start and end of the section.

$$S_i = [ws_{i0}, ws_{i1}], i \in \{1, \dots, n\},$$

$$LP_j = [lp_{j0}, lp_{j1}], j \in \{1, \dots, m\}$$

In the same way, the intersections of those intervals, which we call from now on as “overlap”, is defined by:

$$WS_i \cap LP_j = \begin{cases} [\max(ws_{i0}, lp_{j0}), \min(ws_{i1}, lp_{j1})] & \text{if } (ws_{i0} < lp_{j1}) \wedge (lp_{j0} < ws_{i1}) \\ \emptyset & \text{if not} \end{cases}$$

Thus, the overlap will be well-defined only if the start of one interval occurs before the end of the other.

Then, if we define the length of each interval  $I = [I_0, I_1]$ , as the duration of it, being the null interval of length 0:

$$length(I) = \begin{cases} I_1 - I_0 & \text{if } I \neq \emptyset \\ 0 & \text{if } I = \emptyset \end{cases}$$

We then define the overlap measure:

$$overlap(WS_i, LP_j) = \frac{length(WS_i \cap LP_j)}{length(WS_i)}$$

This is the fraction of  $WS_i$  shared with  $LP_j$ .

We use the WorkSheets components as the denominator since the activities of the instructional syntax are defined around the worksheets.

### 2.3.2 Clustering of the lessons

Now, each class will have potentially a set of  $n \cdot m$  overlaps that can be defined. However, it's to be expected that several of those are empty sets, since the overlap between the first section of one class segmentation and the last section of the other is expected to be empty.

Therefore, for each class  $k \in \{1, \dots, K\}$  we have an overlap matrix  $C_k$ , defined as:

$$C_k = \begin{pmatrix} overlap(WS_1, LP_1) & \dots & overlap(WS_1, LP_m) \\ \vdots & \ddots & \vdots \\ overlap(WS_n, LP_1) & \dots & overlap(WS_n, LP_m) \end{pmatrix}$$

This matrix can be flattened and turned into a vector, and those vectors combined into a new matrix  $C$  of  $K$  rows (one for each class) and  $n \cdot m$  columns (one for each possible overlap).

We applied two clustering algorithms using this matrix. First, we applied the AGNES (Agglomerative Nesting) algorithm (Kaufman and Rousseeuw, 1990). This is a clustering method that starts with each data point as its own group and merges the most similar groups iteratively. This process continues until all elements are grouped together in one cluster. The process is shown as a dendrogram, a tree diagram where branch heights represent similarity.

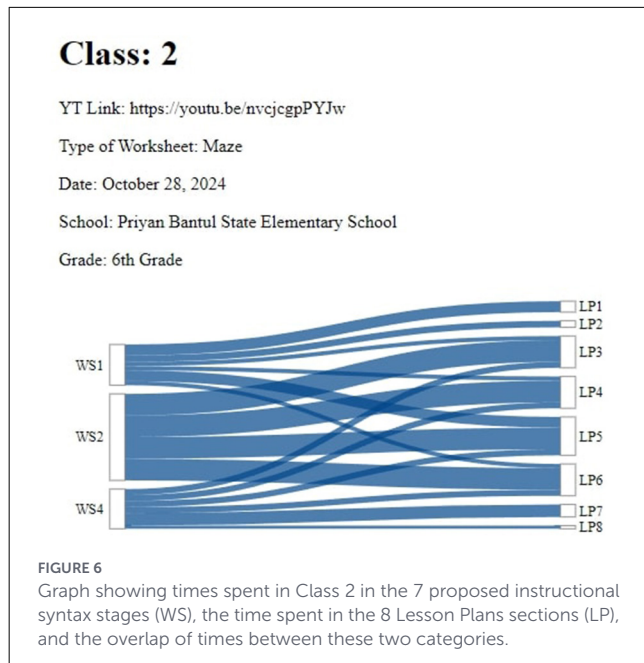
Second, we applied the  $k$ -means clustering algorithm with cluster sizes ranging from 1 to 15.  $K$ -means is a simple machine learning method that groups data into a chosen number of clusters. It works by placing points into the nearest group center, then updating those centers iteratively until the groups reach a stable state. This helps reveal natural patterns or groupings in the data. By plotting the number of clusters vs. the within-cluster sum of squares, we obtained a second clustering of the classes.

Using both AGNES and  $k$ -means provides two complementary views of clustering: AGNES reveals the full hierarchy of group relationships, while  $k$ -means highlights compact clusters around centroids. Combining insights from both methods helps validate results, yielding more reliable groupings and greater confidence in the patterns found in the data. The clustering algorithms can produce slightly different clusters. It's like having two observers grouping clouds. They have slightly different perspectives. However, the commonality across both clusters provides greater robustness.

## 3 Results

After prompt 3, Gemini generated the following 8 sections for Lesson Plans:

- **LP1**—Preliminary Activities (or Introduction);
- **LP2**—Core Activities: Presenting a problem or scenario to students;
- **LP3**—Core Activities: Students engaging with the problem, observing, analyzing, and discussing;
- **LP4**—Core Activities: Working on worksheets (LKPD) individually or in groups;
- **LP5**—Core Activities: Teacher guidance and observation of student progress;



- **LP6**—Core Activities: Presenting and discussing results, with feedback from other groups and teacher reinforcement;
- **LP7**—Closing Activities: Summarizing or reviewing the material learned, and Reflection on the learning experience by students and teachers;
- **LP8**—Closing Activities: Providing information about the next lesson and/or Closing with a prayer and/or greetings.

The eight sections of the lesson plans can be meaningfully grouped into four broader stages that align with the instructional syntax anticipated by one of the authors.

- **Stage I:** Orientation of the problem (LP1, and LP2) frames the lesson by preparing students and introducing a central scenario or problem to spark curiosity.
- **Stage II:** Guiding student investigations (LP3, LP4, LP5) encompasses the phase where students actively engage with the problem through observation, analysis, and worksheet
- **Stage III:** Developing and presenting results (LP6) focuses on group presentations, discussions, and feedback, reinforcing knowledge construction through collaborative sharing and teacher scaffolding.
- **Stage IV:** Evaluating the process and outcomes (LP7, and LP8) emphasizes reflection, review, and closure, helping students consolidate learning and prepare for subsequent lessons.

Together, these stages define a structured instructional syntax that supports problem-solving, problem-posing, and reflective thinking.

We computed the times spent on each stage or section on the vertical axis. The findings are presented in a graph like [Figure 6](#), called a Sankey diagram. Each axis starts at the top (the lesson's initial time) and ends at the bottom. We use one graph for both representations. The time effectively spent on each of the 7 WS

stages according to the proposed instructional syntax is represented on the vertical axis on the right, and the time spent on the LP sections is represented on the vertical axis on the left.

For example, [Figure 6](#) on the left axis shows that Class 2 spent all of the time in WS1, WS2, and WS4. This means that, according to Gemini's automatic analysis of the class video transcription, this class did not spend any time in WS3—Page A: Presentation and Discussion, WS5—Page B: Presentation and Discussion, or WS6—Reflection and Closing. It only spends time in WS1—Prelude, WS2—Page A: Coloring, and WS4—Page B: Coloring. Moreover, the class spent most of the time in WS2. On the other hand, on the right axis, according to Gemini, after reviewing the transcriptions of the class video, [Figure 6](#) shows that Class 2 spent most of the time in LP 3- Students engaging with the problem, LP4- Working on worksheets, LP5- Teacher guidance, and LP6 Presenting and discussing results. However, the alignment of these 4 sections is mainly with coloring page A (WS2).

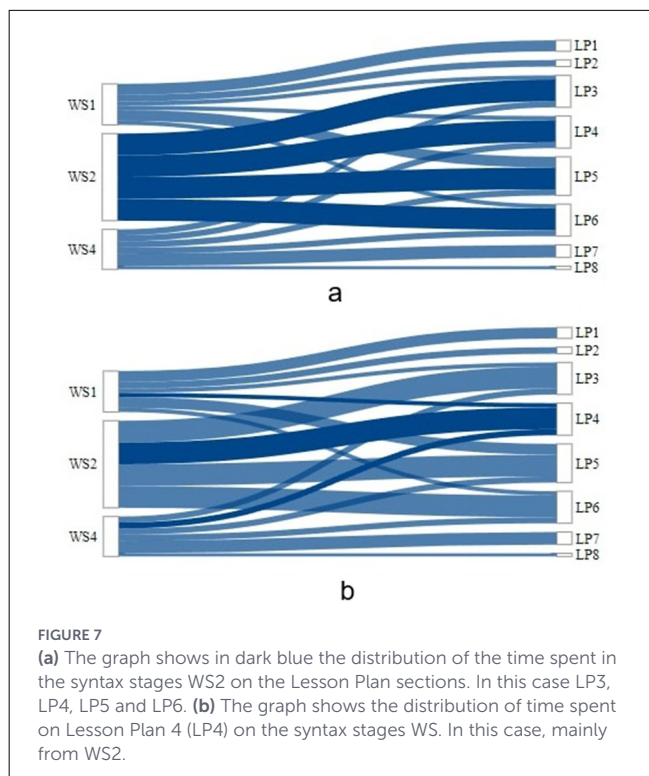
Moreover, placing the mouse cursor in different positions, the graph provides a detailed view of the relationship between the syntax stages and the Lesson Plan sections, showing how classroom time was distributed across both structures. Specifically, it illustrates the time allocated to each syntax stage in each Lesson Plan section and, conversely, how that time was divided among the syntax stages. This one-to-one correspondence offers a precise diagnosis of what actually occurred in the lesson compared to what was both proposed and planned. For example, in Class 2 ([Figure 7](#)), when the mouse cursor is placed on the WS2 syntax stage (left), the graph highlights its overlap with the different Lesson Plans sections. On the other hand, when placing the mouse cursor on the Lesson Plan LP4 (right) the graph highlights its mapping back to the Syntax Stages. The power of these graphs lies in their ability to reveal, with clarity and detail, how teaching unfolded in practice, significantly enhancing our understanding of classroom implementation.

After running the first clustering algorithm, the Agglomerative Nesting algorithm, we obtained the dendrogram shown in [Figure 8](#). This is a tree-like diagram that visually represents a hierarchy of clusters. It shows that classes 1 and 3 are very similar, in terms of the time spent on the WS instructional syntax stages, the time spent on the lesson plans sections, and the corresponding overlaps. This is illustrated in [Figure 9](#). It shows that these classes spent most of their time in WS1, WS2, WS4, and WS6, with the majority in WS2. Furthermore, the distribution of LPs is similar, with both classes following a sequential pattern from LP1 to LP8, unlike in other classes. Furthermore, the correspondence between WS2 and LPs, as well as WS3 and LPs, is similar in both classes.

From [Figure 8](#), we see that classes such as class 10 are very similar to classes 0A and 0B, which were taught by experienced teachers in this WS using A/B coloring books. This is surprising, given the very short training time for teachers, and indicates that it is a strategy that is not difficult to adopt and implement in classes. If we want to define only four clusters from the dendrogram, then the four clusters are defined by the four colors in [Figure 9](#).

The first cluster is formed by classes 1, 2, 7, 2, 4, 12, 8, 16, 17, 8, 11, 14, 20, and 16. Cluster 2 contains only class 5, cluster 3 contains only class 15, and cluster 4 contains class 10 and the two expert classes 0B and 0A.

[Figure 10](#) shows that the times dedicated to the proposed instructional syntax stages and to the time dedicated to the lesson



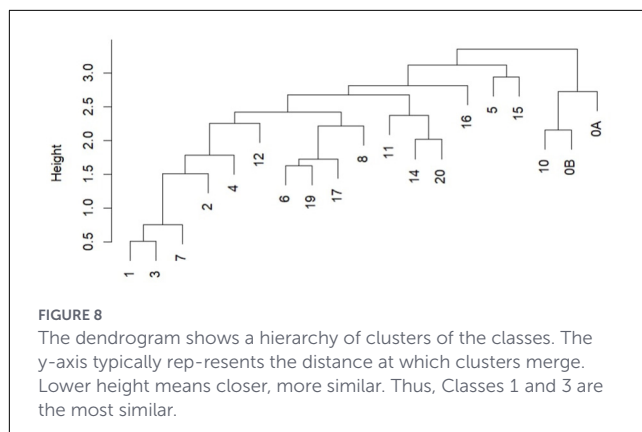
plan sections are similar, as well as the overlaps between stages and sections in these three classes. It is interesting to note that the time spent on WS4 is reflected in the number of page B worksheets with students' work. For example, from class 10, we collected 9 pages.

Then we applied the second clustering algorithm, *k*-means. We compute clusters using *k*-means for clusters of size 1 to 15, and by plotting the number of clusters vs. the within-cluster sum of squares (Figure 11). It is desirable to have a small number of clusters. However, if we cluster classes into a small number, the difference between classes within the clusters may be too high. If we consider more clusters, the differences between classes within the clusters diminish.

We chose 4 clusters for our clustering because, beyond that number, we observe a slight but significant decrease in the slope of the curve in Figure 11, indicating that the benefit of using a larger number of clusters is significantly less each time. The results shown in Figure 12 differ slightly from those obtained using this method. However, classes 1 and 3 are grouped together again.

### 4 Discussion

The core of the proposed instructional syntax is structured metacognitive reflection at three levels (individual, peer, and collective). This directly addresses the need to enhance students' metacognitive skills and their mathematical and computational thinking in an integrated way. This study provides compelling evidence of both the feasibility and promise of introducing the novel A/B Coloring Book instructional syntax in Indonesian elementary schools. A total of 17 pilot lessons conducted in 17 schools with 215 students generated an exceptionally rich

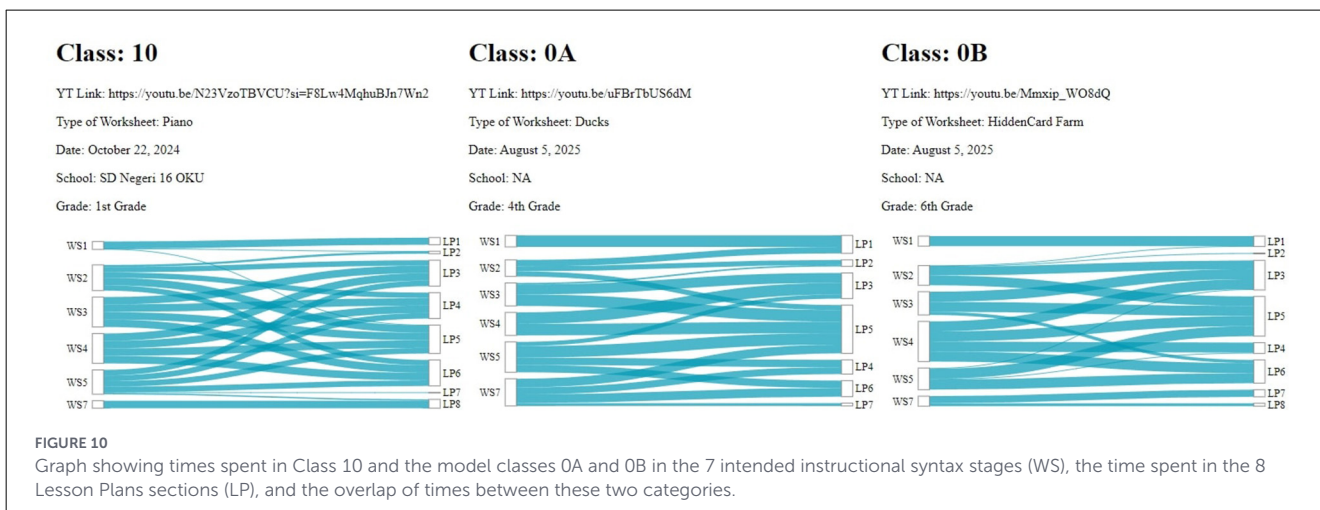
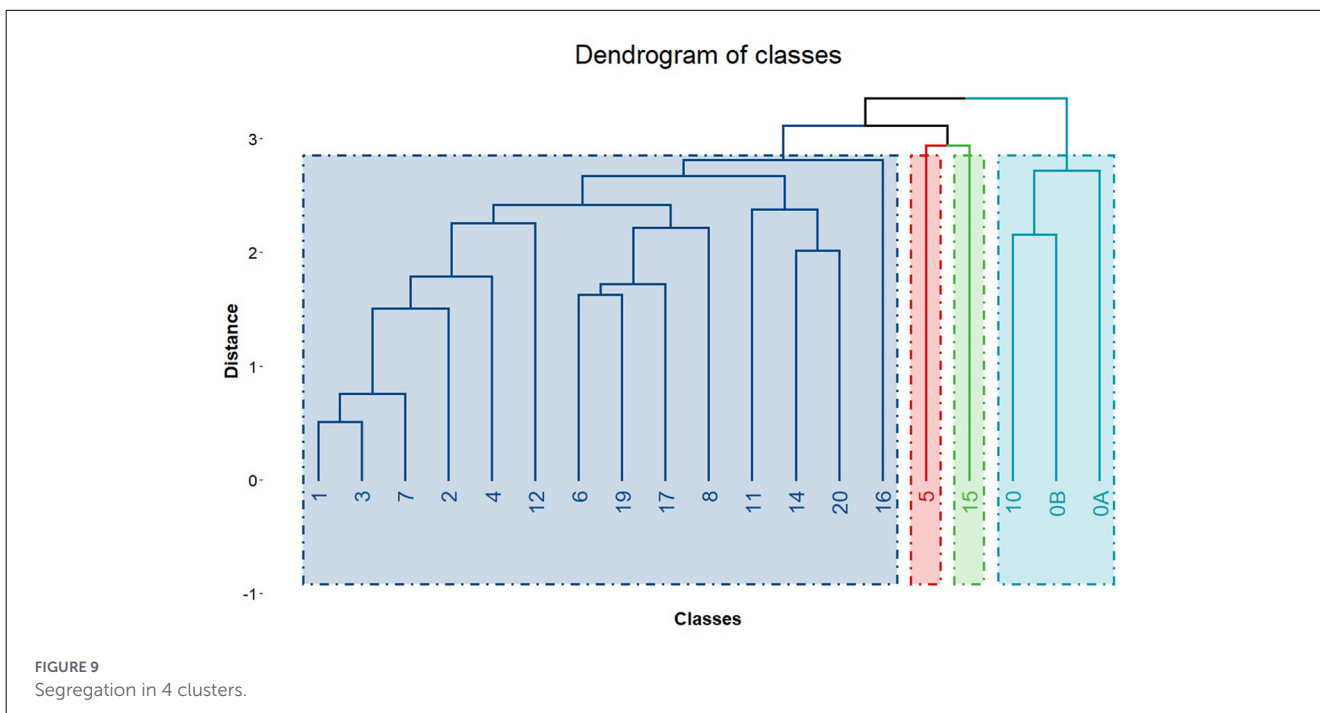


dataset, including detailed lesson plans, video recordings, and student worksheets. These authentic materials allowed us to analyze classroom practices with high granularity. The diversity of schools, grade levels, and teaching styles strengthens the external validity of our findings, demonstrating that the syntax can be implemented under real-world conditions.

The study successfully operationalized a wide array of CT components through the A/B coloring worksheets, which served as the primary vehicle for student engagement and deep understanding. These components span a cognitive spectrum from basic whole numbers and fractional thinking to more advanced logical, recursive, inferential, and algorithmic reasoning. As detailed in Table 1, students engaged in tasks such as fractional reasoning, spatial relations, and recursive containment, as well as statistical and probabilistic thinking (e.g., identifying likely outcomes based on cues) and causal thinking (e.g., executing rules based on relative spatial positions). Notably, the activities incorporated both mathematical and computational modeling, requiring students to simulate biological behaviors—such as worms or “Pac-Man-like” agents navigating boards via gradient descent. This approach bridges the fundamental abstractions of Turing machine thinking, such as algorithmic and recursive thinking, to von Neumann thinking with cellular automata, allowing primary students to internalize complex logic through active, embodied, and visual execution.

The inclusion of agent-based modeling, following the von Neumann framework (Agüera y Arcas, 2025a; Araya, 2021b, 2022), represents a powerful component of CT explicitly aligned with mathematical and computational modeling (Table 1). By modeling organisms as digital agents performing simple local computations that generate emergent behaviors, students acquire the conceptual tools and way of thinking to develop a deep understanding of complex phenomena in biological and social systems. The LLM-supported transcript analysis confirmed that this instructional syntax—which integrates written argumentation, problem posing, prompting, and metacognition—creates a traceable record of classroom implementation. Results indicate that even sophisticated CT components, such as recursive reasoning and agent-based computational models, can be effectively integrated into primary mathematics through this reflective instructional syntax.

Regarding RQ1, our data confirm that the computationally integrated syntax is implementable even after minimal training.



While adoption varied, several teachers reproduced sequences very similar to those modeled by expert practitioners, suggesting that the syntax is intuitive and adaptable. Others showed partial implementation, particularly by omitting reflective components, underscoring areas for future professional development. Nevertheless, the overall consistency of implementation across diverse contexts supports the claim that this syntax can be applied broadly.

Regarding RQ2, our AI-based analysis, powered by Gemini, allowed us to detect precise patterns of overlap between the instructional syntax stages and the sections of the teachers' lesson plans. This capacity to compute and visualize alignment is a methodological breakthrough. Figures such as Figure 7 exemplify this novelty: on the left, the distribution of time from the syntax stage WS2 across multiple Lesson Plan sections is shown, while on the right, the distribution of Lesson Plan 4 (LP4) is traced

back across the syntax stages. This novel dual representation makes visible the one-to-one correspondences between planned structures and enacted classroom practices. It reveals exactly how teachers interpreted the proposed syntax, where they converged with or diverged from it, and how much time was invested in each stage.

These findings directly address the research questions by demonstrating both the feasibility of implementing the proposed syntax and the utility of AI in identifying actionable patterns to improve CT instruction.

To the best of our knowledge, such visual information is a new device to present what is happening in classroom practices. This type of graph is not only descriptive but also diagnostic, as it pinpoints how a class unfolded in practice compared to both the teacher's plan and the proposed syntax. It has great potential for teacher training. By showing

teachers precisely how their timing and sequencing map onto an intended model, trainers can provide targeted, evidence-based feedback. Furthermore, the clustering analysis revealed that even with limited preparation, some teachers' classes resembled expert implementations, highlighting the approach's accessibility.

The integration of AI into classroom analysis enables the precise identification of instructional patterns and gaps, offering teachers actionable insights to improve computational thinking

instruction. By transforming classroom discourse into visual representations, AI facilitates intuitive understanding of teaching dynamics. These visualizations can help educators reflect on their practice, recognize missed opportunities for dialogic engagement, and adjust lesson flow accordingly. Moreover, this AI-driven feedback can support formative assessment and professional development by highlighting strengths and areas for growth. This approach can empower teachers to refine their pedagogical strategies, fostering deeper student reasoning and more effective implementation of computational thinking frameworks.

The joint application of Large Language Models (LLMs) with unsupervised clustering algorithms (Altamirano et al., 2022) to classroom transcriptions offers a paradigm shift in educational analysis. This computational approach provides novel tools for understanding the actual mechanics of classroom instruction. By analyzing dialogue at a scale and high degree of granularity, the methodology functions as a microscope for exploring the "dark matter" of the classroom. These are the subtle, high-frequency, and often overlooked patterns of interaction that drive pedagogical efficacy. This innovative combination allows researchers to move beyond surface-level observations to identify core mechanisms of teaching.

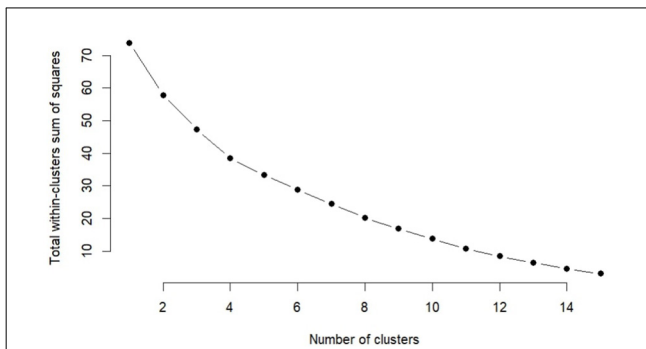


FIGURE 11 Sum of the squares of the differences within clusters as a function of the number of clusters.

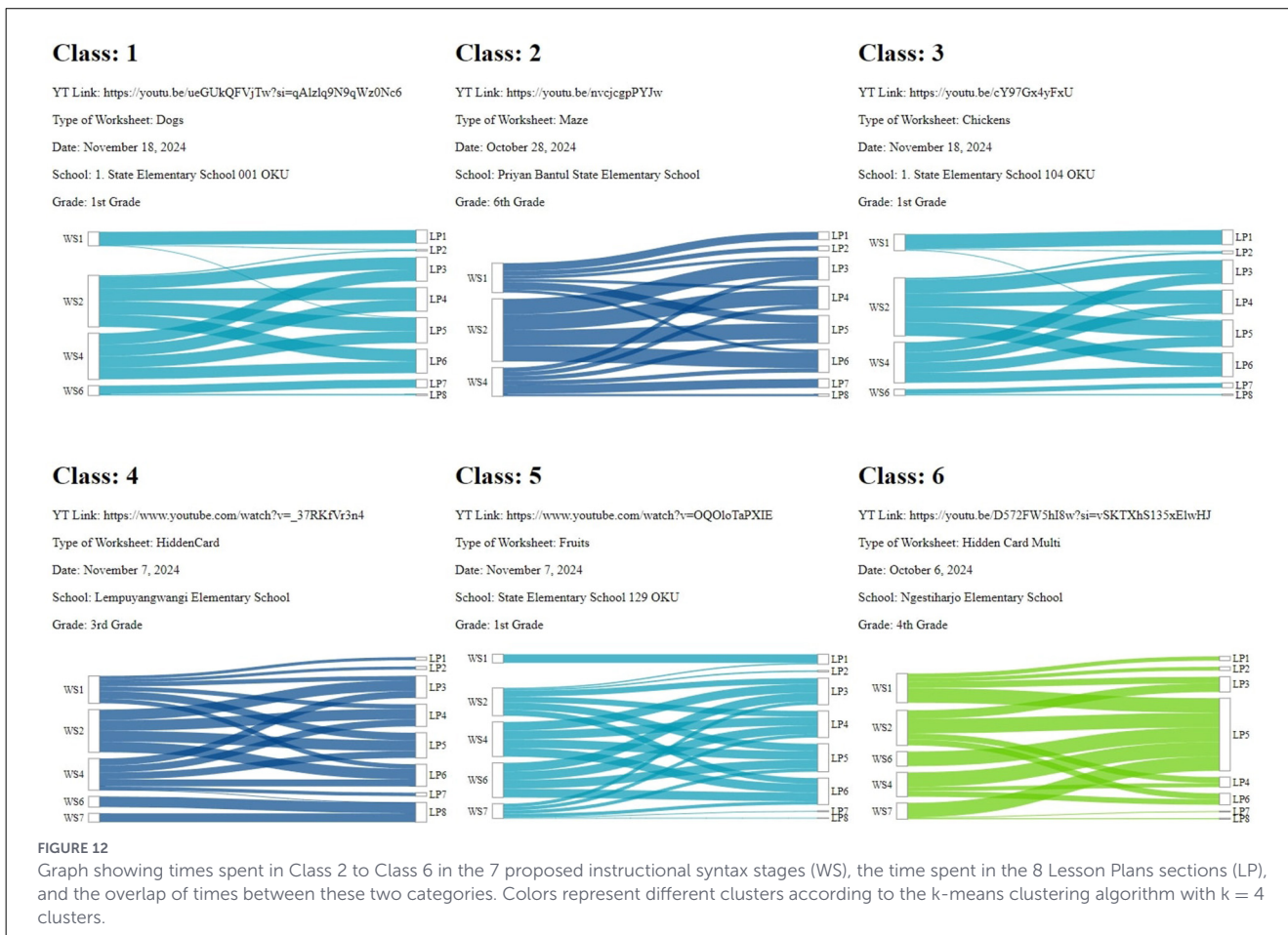


FIGURE 12 Graph showing times spent in Class 2 to Class 6 in the 7 proposed instructional syntax stages (WS), the time spent in the 8 Lesson Plans sections (LP), and the overlap of times between these two categories. Colors represent different clusters according to the k-means clustering algorithm with  $k = 4$  clusters.

The broader contribution of this work lies in combining a novel pedagogy, an instructional syntax that fosters reflection both individually and in collaboration with peers and the whole class, with AI-powered analysis of authentic classroom data. By applying Gemini Reasoner to full class transcripts, we reduce the risks of bias, inconsistency, fatigue, or subjective interpretation inherent in human coding. This approach makes the analysis scalable, replicable, and more objective. It also allows researchers to identify fine-grained patterns of practice across large datasets, opening the door to comparative studies across schools, regions, and even countries.

However, this study has several important limitations that are potential sources of error and noise. First, it relies on transcription of classroom sessions. As a result, a significant amount of visual and auditory information is missing, limiting our ability to capture precisely what actually occurred in the classroom. Second, the transcripts do not include diarization, meaning they do not distinguish between speakers. This constrains the analysis of dialogues and particularly the contributions made by students. Third, LLMs themselves still have limitations in fully understanding the nature of classroom transcripts. Although their comprehension is improving rapidly, errors persist, and results are highly dependent on the quality of the prompts. Future work will focus on improving prompt engineering to enhance accuracy. Fourth, the video recordings were collected using teachers' smartphones, a practical, low-cost, and independent approach that requires no additional personnel. This is critical because specialized audio technical teams are not available in schools. However, the audio quality is not ideal, affecting the accuracy of the transcription. Despite these limitations, our method already provides an automatic, highly granular description of classroom activity, offering a valuable input for assessing feasibility and identifying critical aspects of practice that can be improved.

Future work should build upon this initial study by expanding the number of classes and schools, thereby deepening both statistical and qualitative insights. Equally crucial is our parallel effort to analyze student worksheets, which capture reasoning, creativity, and clarity of problem posing. Linking teacher implementation data to student outcomes will enable us to assess the comprehensive pedagogical impact of the A/B syntax. A larger sample will allow us to examine variability more comprehensively and to refine the syntax based on diverse contexts of practice. Equally important is our ongoing study of students' worksheets, which capture not only their coloring but also their written explanations and the problems they pose for their peers. This data offers a unique window into children's reasoning processes, creativity, and clarity of formulation. In particular, we are examining how problem posing evolves as students generate tasks for classmates and how peers can interpret and respond effectively. This analysis will allow us to determine whether the method enhances precision, expressiveness, and creativity in mathematical and computational thinking.

As articulated in our research questions, this study critically examines the feasibility of implementing the proposed instructional syntax for teaching CT. This investigation is non-trivial, as the syntax necessitates a fundamental shift in pedagogical strategies,

moving away from traditional instruction toward structured, multi-level metacognitive reflection. While the current work focuses on classroom dynamics and student output, a subsequent study in preparation specifically addresses teacher perceptions to better understand the professional transition required. Looking forward, we intend to conduct a Randomized Controlled Trial (RCT) to rigorously quantify effect sizes across the diverse CT components identified, using third parties' performance measures (Duflo et al., 2007; Pellegrini et al., 2021; Cheung and Slavin, 2016). This experimental design will follow established RCT methodology in mathematics education, as we have already successfully measured learning gains following a year-long intervention in primary mathematics education (Araya et al., 2025b) and another semester-long intervention comprising 1 or 2 weekly mathematics activities (Araya and Diaz, 2020). A cornerstone of such rigorous evaluation is the precise measurement of implementation fidelity. Without high fidelity, determining true effect sizes becomes impossible, as execution variance can mask the intervention's efficacy. Therefore, this study lays the groundwork by utilizing LLM-supported analysis to establish a baseline for how faithfully the syntax is enacted in authentic primary settings.

In sum, this study answers both research questions affirmatively. The syntax can be applied in real classrooms, and AI-based tools can effectively identify some of the patterns that matter for student learning. At the same time, it introduces a methodological innovation—visual overlap graphs—that offers a level of transparency into classroom practice never before achieved. These contributions position the A/B Coloring Book syntax, combined with AI-supported analysis, as a powerful, scalable strategy for advancing CT and mathematics instruction worldwide.

Despite the considerable challenge of integrating mathematical thinking and computational thinking in primary education, our study found that it is feasible to implement classroom instruction that simultaneously integrates these domains. Moreover, the instructional syntax we propose can help teach, reflect, and develop an initial understanding of fundamental computational thinking concepts that support connections across core computational and biological, social, and psychological concepts, such as cellular automata, spatial navigation, and recursion. The pilot implementation shows that young learners can engage with complex reasoning and that reflective, dialogic instruction can bridge disciplinary boundaries, preparing students for integrative, real-world problem-solving and problem-posing.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

The studies involving humans were approved by Ethics Committee of SEAMEO Regional Center for QITEP in Mathematics. The studies were conducted in accordance with the local legislation and institutional requirements. Written

informed consent for participation in this study was provided by the participants' legal guardians/next of kin. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

## Author contributions

SD: Conceptualization, Validation, Supervision, Project administration, Funding acquisition, Writing – original draft, Formal analysis, Resources. WW: Writing – original draft, Validation. OU: Software, Writing – original draft, Formal analysis, Visualization. RR: Writing – original draft, Validation. AS: Writing – original draft, Validation. RA: Writing – original draft, Investigation, Supervision, Writing – review & editing, Methodology, Funding acquisition, Conceptualization, Resources, Validation, Project administration, Formal analysis.

## Funding

The author(s) declared that financial support was received for this work and/or its publication. The research has received funding support from the SEAMEO Regional Centre for QITEP in Mathematics Decision Letter 650.1/QiM.1.1/SK.2/2024 dated September 25, 2024 DIPA BBGP D.I. Yogyakarta, and ANID CIAE CIA250005 and ANID Exploración 13240075.

## Acknowledgments

SEAMEO Regional Mathematics QITEP Centre Decision Letter 650.1/QiM.1.1/SK.2/2024 dated September 25, 2024

## References

- Aghion, P., and Bunel, S. (2024). *AI and Growth: Where Do We Stand?* Available online at: <https://www.frbsf.org/wp-content/uploads/AI-and-Growth-Aghion-Bunel.pdf> (Accessed April 8, 2024).
- Agüera y Arcas (2025b). What is the future of intelligence? The answer could lie in the story of its evolution. *Nature*. 647, 846–850 doi: 10.1038/d41586-025-03857-0
- Agüera y Arcas, B. (2025a). *What is Intelligence? Lessons from AI about Evolution, Computing, and Minds*. MIT Press.
- Alexander, R. (2020). *A Dialogic Teaching Companion, 1st Edn*. London: Taylor & Francis Group. doi: 10.4324/9781351040143
- Altamirano, M., Uribe, P., Schlotterbeck, D., Jiménez, A., Araya, R., van der Molen Moris, J. et al. (2022). Unsupervised characterization of lessons according to temporal patterns of teacher talk via topic modeling. *Neurocomputing* 484, 211–222. doi: 10.1016/j.neucom.2021.09.078
- Araya, R. (2021a). “Gamification strategies to teach algorithmic thinking to first graders,” in *Advances in Human Factors in Training, Education, and Learning Sciences*, eds. S. Nazir, T. Z. Ahram, and W. Karwowski (Springer International Publishing), 133–141.
- Araya, R. (2021b). Enriching elementary school mathematical learning with the steepest descent algorithm. *Mathematics* 9:1197. doi: 10.3390/math9111197
- Araya, R. (2022). “Is it feasible to teach agent-based computational modeling to elementary and middle school students?” in *Proceedings of the Singapore National*

DIPA BBGP D.I. Yogyakarta, and Funding from ANID CIAE CIA250005 and ANID Exploración 13240075 are gratefully acknowledged.

## Conflict of interest

The author(s) declared that that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. We have used Gemini and GPT5 to review sentences. and correct style.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Academy of Science*. Available online at: <https://www.worldscientific.com> (Accessed December 2, 2022).

Araya, R. (2023). What and how to teach mathematics for the future? *Math. Educ.* 4, 84–108. Available online at: <https://ame.org.sg/tme2023-vol-4-no-2-pp-84-108/> (Accessed January 02, 2026).

Araya, R. (2024). “AI as a co-teacher: enhancing creative thinking in underserved areas,” in *International Conference on Computers in Education*. doi: 10.58459/icce.2024.5048

Araya, R. (2025a). Do chains-of-thoughts of large language models suffer from hallucinations, cognitive biases, or phobias in bayesian reasoning? (version 1). *arXiv*.

Araya, R. (2025b). *LLMs' inner voice illusions and disagreements can be converted into a new tool for teaching deductive reasoning* [manuscript submitted for publication].

Araya, R. (2025c). A/B coloring textbooks: the next orbis sensualium pictus? *J. Southeast Asian Educ.* [Epub ahead of print].

Araya, R., Aguirre, C., and Díaz, M. (2025a). Effect of embodied argumentation with hand puppets in fourth-graders' mathematical thinking. *IxD&A* 63, 91–107. doi: 10.55612/s-5002-063-005

Araya, R., Arias, E., Botton, N., and Cristia, J. (2025b). Integrating learning platforms within regular school time: experimental evidence from Chilean primary schools. *Econ. Educ. Res.* 106:102647. doi: 10.1016/j.econedurev.2025.102647

- Araya, R., and Diaz, K. (2020). Implementing government elementary math exercises online: positive effects found in rct under social turmoil in Chile. *Educ. Sci.* 10:244. doi: 10.3390/educsci10090244
- Araya, R., and González, P. (2025). Determinants of trust: evidence from elementary school classrooms. *J. Intell.* 13:165. doi: 10.3390/jintelligence13120165
- Araya, R., and Isoda, M. (2023). Unplugged computational thinking with colouring books. *J. Southeast Asian Educ.* 1, 72–91.
- Asian Development Bank (2022). *Technology-Enabled Innovation in Education in Southeast Asia (TIESEA) Diagnostic Assessment Report – Indonesia Country Report (Diagnostic Assessment Report) [Indonesia Country Report]*. Asian Development Bank. Available online at: [https://tiesea.org/wp-content/uploads/2022/05/Diagnostic-Assessment-Report-Indonesia\\_TIESEA.pdf](https://tiesea.org/wp-content/uploads/2022/05/Diagnostic-Assessment-Report-Indonesia_TIESEA.pdf) (Accessed May 11, 2022).
- Bengio, Y. (2004). *AI can learn to think before it speaks*. Available online at: <https://www.ft.com/content/894669d6-d69d-4515-a18f-569afb710e8> (Accessed March 16, 2025).
- Caballero, D., Araya, R., Kronholm, H., Viiri, J., Mansikkaniemi, A., Lehesvuori, S., et al. (2017). “ASR in classroom today: automatic visualization of conceptual network in science classrooms,” in *Data Driven Approaches in Digital Education*, eds. É. Lavoué, H. Drachsler, K. Verbert, J. Broisin, and M. Pérez-Sanagustin (Cham: Springer International Publishing), 541–544.
- Changri, N., Inprasitha, M., Araya, R., and Isoda, M. (2025). “Developing logical thinking” in *Early Mathematics Through Colouring Activity of the 48th Conference of the International Group for the Psychology of Mathematics* (Santiago: PME).
- Cheung, A. C. K., and Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educ. Res.* 45, 283–292. doi: 10.3102/0013189X16656615
- Danoebroto, S. W., Suyata, and Jailani. (2024). Teachers’ efforts to promote students’ mathematical thinking using the ethnomathematics approach. *Math. Teach. Res. J.* 16, 207–231.
- Denning, P. (2017). Computational thinking in science. *Am. Sci.* 105:1. doi: 10.1511/2017.124.13
- Didolkar, A., Ballas, N., Arora, S., and Goyal, A. (2025). Metacognitive reuse: turning recurring LLM reasoning into concise behaviors (No. arXiv:2509.13237). arXiv.
- Dini, I., Kim, S., and Nomura, S. (2024). *Teacher Practices in Indonesia: Results of the Teach Primary Classroom Observation Study*. World Bank Group. Available online at: <http://documents.worldbank.org/curated/en/099090924033026750> (Accessed September 9, 2024).
- Duflo, E., Glennerster, R., and Kremer, M. (2007). Using randomization in development economics research: a toolkit. *Handb. Dev. Econ.* 4, 3895–3962. doi: 10.1016/S1573-4471(07)04061-2
- Fabiano, F., Ganapini, M. B., Loreggia, A., Mattei, N., Murugesan, K., Pallagani, V., et al. (2025). Thinking fast and slow in human and machine intelligence. *Commun. ACM* 68, 72–79. doi: 10.1145/3715709
- Fajri, M., and Yurniawati, Y. (2019). Computational thinking, mathematical thinking berorientasi gaya kognitif pada pembelajaran matematika di sekolah dasar. *Dinamika J. Ilm. Pendidik. Dasar* 1, 1–18.
- Fauzi, A. L., Kusumah, Y. S., Nurlaelah, E., and Juandi, D. (2024). Computational thinking in mathematics education: a systematic literature review on its implementation and impact on students’ learning. *J. Kependidikan* 10:640. doi: 10.33394/jk.v10i2.11140
- Fitriyah, Y., Wahyudin, W., Nurhayati, H., and Febrianti, T. S. (2024). Indonesian students’ computational thinking performance based on level and gender. *Int. J. Pedagog. Teach. Educ.* 8:50. doi: 10.20961/ijpte.v8i1.89464
- GEM Report UNESCO (2023). *Technology in education: a case study on Indonesia*. GEM Report UNESCO. doi: 10.54676/WJMY7427
- Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., et al. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* 645, 633–638. doi: 10.1038/s41586-025-09422-z
- Iordanou, K., and Rapanta, C. (2021). Argue with me: a method for developing argument skills. *Front. Psychol.* 12:631203. doi: 10.3389/fpsyg.2021.631203
- Ippolito, D., and Zhang, Y. (2025). AI can learn to show its workings through trial and error. *Nature* 645, 594–595. doi: 10.1038/d41586-025-02703-7
- Joseph, J. (2025). The algorithmic self: how AI is reshaping human identity, introspection, and agency. *Front. Psychol.* 16:1645795. doi: 10.3389/fpsyg.2025.1645795
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, New York, NY: Straus and Giroux.
- Kargupta, P., Li, S. S., Wang, H., Lee, J., Chen, S., Ahia, O., et al. (2025). Cognitive foundations for reasoning and their manifestation in LLMs. *ArXiv*. Available online at: <https://arxiv.org/abs/2511.16660> (Accessed November 20, 2025).
- Kaufman, L., and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis 1st Edn*. New York, NY: Wiley.
- Kazlaris, I., Antoniou, E., Diamantaras, K., and Bratsas, C. (2025). From illusion to insight: a taxonomic survey of hallucination mitigation techniques in LLMs. *AI* 6:260. doi: 10.3390/ai6100260
- Kross, E. (2021). *Chatter: The voice in our head, why it matters, and how to harness it*. New York, NY: Crown.
- Kuhn, D. (2015). Thinking together and alone. *Educ. Res.* 44, 46–53. doi: 10.3102/0013189X15569530
- Lehesvuori, S., and Ametller, J. (2021). Exploring coherence and authorship in pedagogical link-making in science. *Int. J. Sci. Educ.* 43, 2791–2813. doi: 10.1080/09500693.2021.1991599
- Lehesvuori, S., Kelly, S., and Ramnarain, U. (2023). “Responsible talk in science,” in *Science Teacher Learning for the 21st Century And Beyond 1st Edn*. eds. N. Petersen, R. Umesh, D. Kruger, L. Mavuru, and A. Lubbe (Hatfield; Van Schaik Publishers), 43–58.
- Lehesvuori, S., Urrutia, F., Heilala, V., Araya, R., and Hämäläinen, R. (2025). Discovering technology-aided possibilities for automatic analysis of science teacher questions. *Educ. Technol. Res. Dev.* 73, 1325–1345. doi: 10.1007/s11423-025-10545-3
- Lehesvuori, S., Viiri, J., Rasku-Puttonen, H., Moate, J., and Helaakoski, J. (2013). Visualizing communication structures in science classrooms: tracing cumulativity in teacher-led whole class discussions. *J. Res. Sci. Teach.* 50, 912–939. doi: 10.1002/tea.21100
- Moser, J. S., Dougherty, A., Mattson, W. I., Katz, B., Moran, T. P., Guevarra, D., et al. (2017). Third-person self-talk facilitates emotion regulation without engaging cognitive control: converging evidence from ERP and fMRI. *Sci. Rep.* 7:4519. doi: 10.1038/s41598-017-04047-3
- O’Connor, S. (2020). The natural selection of private and inner speech. *Front. Psychol.* 11:163. doi: 10.3389/fpsyg.2020.00163
- OECD (2023). *PISA 2022 Results (Volume I and II)—Country Notes: Indonesia*. Available online at: [https://www.oecd.org/en/publications/pisa-2022-results-volume-i-and-ii-country-notes\\_ed6fbcc5-en/indonesia\\_c2e1ae0e-en.html](https://www.oecd.org/en/publications/pisa-2022-results-volume-i-and-ii-country-notes_ed6fbcc5-en/indonesia_c2e1ae0e-en.html) (Accessed December 5, 2023).
- OECD (2024). *PISA Results 2022 (Volume III)—Factsheets: Indonesia*. OECD Publishing. Available online at: [https://www.oecd.org/en/publications/2024/06/pisa-2022-results-volume-iii-country-notes\\_72b418f8/indonesia\\_cf276198.html](https://www.oecd.org/en/publications/2024/06/pisa-2022-results-volume-iii-country-notes_72b418f8/indonesia_cf276198.html) (Accessed June 18, 2024).
- Pellegrini, M., Neitzel, A., Lake, C., and Slavin, R. (2021). Effective programs in elementary mathematics: a best-evidence synthesis. *AERA Open* 7, 1–29. doi: 10.1177/2332858420986211
- Permana, R., Fitriani, D., and Ramadham, H. (2022). Penerapan berpikir komputasional dalam pembelajaran matematika sekolah dasar. *J. Pendidik. Mat. Indones.* 7, 101–112.
- Prahmana, R. C. I., Kusaka, S., Peni, N. R. N., Endo, H., Azhari, A., and Tanikawa, K. (2024). Cross-cultural insights on computational thinking in geometry: Indonesian and Japanese students’ perspectives. *JME* 15, 613–638. doi: 10.22342/jme.v15i2.pp613-638
- Roby, A. C., and Kidd, E. (2008). The referential communication skills of children with imaginary companions. *Dev. Sci.* 11, 531–540. doi: 10.1111/j.1467-7687.2008.00699.x
- Rosali, D. F., and Suryadi, D. (2021). An analysis of students’ computational thinking skills on the number patterns lesson during the Covid-19 pandemic. *Formatif: J. Ilm. Pendidik. MIPA* 11, 217–232. doi: 10.30998/formatif.v11i2.9905
- Schank, P., Jenks, L., Barth, C., Crawford, C., Powers, K., and Fusco, J. (2025). *Classroom Discourse*. Washington: Digital Promise.
- Schlotterbeck, D., Araya, R., Caballero, D., Jimenez, A., Lehesvuori, S., and Viiri, J. (2020). Assessing teacher’s discourse effect on students’ learning: a keyword centrality approach,” in *Addressing Global Challenges and Quality Education. EC-TEL 2020. Lecture Notes in Computer Science, vol 12315*, eds. C. Alario-Hoyos, M. Rodríguez-Triana, M. Scheffel, I. Arnedillo-Sánchez, and S. Dennerlein (Cham: Springer). doi: 10.1007/978-3-030-57717-9\_8
- SMERU Research Institute (2022). “Diagnostic report digital skills landscape in Indonesia. the smeru research institute,” in *Partnership with Digital Pathways at University of Oxford and the United Nations Economic and Social Commission for Asia and the Pacific (ESCAP)* (Jakarta: The SMERU Research Institute), 1–74. Available online at: <https://smeru.or.id/en/publication/diagnostic-report-digital-skills-landscape-indonesia> (Accessed March 4, 2026).
- Somsaman, K., Isoda, M., Asami, and Araya, R. (2024). *Guidebook for Unplugged Computational Thinking*. Bangkok: The Southeast Asian Ministers of Education Organization Regional Centre for STEM Education (SEAMEO STEM-ED).
- Stanovich, K. E., and Toplak, M. E. (2023). Actively open-minded thinking and its measurement. *J. Intell.* 11:27. doi: 10.3390/jintelligence11020027
- Stanovich, K. E., West, R. F., and Toplak, M. E. (2016). *The Rationality Quotient: Toward a Test of Rational Thinking*. Cambridge: MIT Press.
- Street, W., Siy, J. O., Keeling, G., Baranes, A., Barnett, B., McKibben, M., et al. (2024). LLMs achieve adult human performance on higher-order theory of mind tasks. *ArXiv*. Available online at: <https://arxiv.org/abs/2405.18870> (Accessed February 15, 2026).
- Suarsana, I. M., Dasari, D., and Nurlaelah, E. (2024). “Integration of computational thinking in mathematics education in Indonesia,” in *Proceedings of the 4th International Conference on Education and Technology (ICETECH 2023)* eds. J.

- Handhika, M. Lukitasari, S. Ricahyono, and D. A. Nugraha (Pairs: Atlantis Press International), 211–226
- Tapia-Mandiola, S., and Araya, R. (2024). From play to understanding: large language models in logic and spatial reasoning coloring activities for children. *AI* 5, 1870–1892. doi: 10.3390/ai5040093
- Tedre, M., and Denning, P. (2016). *The Long Quest for Computational Thinking*. Available online at: <http://denninginstitute.com/pjd/PUBS/long-quest-ct.pdf> (Accessed November 24, 2016).
- Tigros (n.d.). *Whisperer [Computer software]*. Available online at: <https://github.com/tigros/Whisperer> (Accessed March 4, 2026).
- UNICEF (2021). *Situation Analysis on Digital Learning in Indonesia*. Available online At: <https://www.unicef.org/indonesia/media/8766/file/DigitalLearningLandscapeinIndonesia.pdf> (Accessed February 1, 2021).
- Urrutia, F., and Araya, R. (2024). Who's the best detective? Large Language Models vs. Traditional Machine Learning in Detecting Incoherent Fourth Grade Math Answers. *J. Educ. Comput. Res.* 61, 1723–1754. doi: 10.1177/07356331231191174
- Vygotsky, L. S. (1978). *Mind in Society: Development of Higher Psychological Processes*. eds. M. Cole, V. Jolm-Steiner, S. Scribner, and E. Souberman (Cambridge, MA: Harvard University Press).
- Wing, J. (2006). Computational Thinking. *Commun. AC* 49, 33–35. doi: 10.1145/1118178.1118215
- Xu, S., Huang, X., Lo, C. K., Chen, G., and Jong, M. S. (2024). Evaluating the performance of ChatGPT and GPT-4o in coding classroom discourse data: a study of synchronous online mathematics instruction. *Comput. Educ. AI* 7:100325. doi: 10.1016/j.caeai.2024.100325