Computers in Human Behavior Reports 15 (2024) 100466

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Full length article

# Development of algorithmic thinking skills in K-12 education: A comparative study of unplugged and digital assessment instruments

Giorgia Adorni [a,*], Igor Artico [b], Alberto Piatti [c], Elia Lutz [d], Luca Maria Gambardella [a], Lucio Negrini [c,e], Francesco Mondada [f,g], Dorit Assaf [d]

[a] Dalle Molle Institute for Artificial Intelligence (IDSIA), Università della Svizzera Italiana and University of Applied Sciences and Arts of Southern Switzerland (USI-SUPSI), Polo universitario Lugano - Campus Est, Via la Santa 1, CH-6962, Lugano-Viganello, Switzerland
[b] Università della Svizzera Italiana, Polo universitario Lugano - Campus Est, Via la Santa 1, CH-6962, Lugano-Viganello, Switzerland
[c] Department of Education and Learning (DFA), University of Applied Sciences and Arts of Southern Switzerland (SUPSI), Piazza S. Francesco 19, CH-6600, Locarno, Switzerland
[d] School of Education, University of Applied Sciences and Arts Northwestern Switzerland (FHNW), Bahnhofstrasse 6, CH-5210, Windisch, Switzerland
[e] Laboratory media and MINT, Department of Education and Learning (DFA), University of Applied Sciences and Arts of Southern Switzerland (SUPSI), Piazza S. Francesco 19, CH-6600, Locarno, Switzerland
[f] Mobile Robotic Systems Group (MOBOTS), Ecole Polytechnique Fédérale de Lausanne (EPFL), Station 9, CH-1015, Lausanne, Switzerland
[g] Center for Learning Sciences (LEARN), Ecole Polytechnique Fédérale de Lausanne (EPFL), Station 9, CH-1015, Lausanne, Switzerland

## ARTICLE INFO

*Keywords:*
Teaching/learning strategies
21st century abilities
Evaluation methodologies
Elementary education
Secondary education

## ABSTRACT

In the rapidly evolving landscape of digital competencies, the need for a robust and universal method to assess students' algorithmic thinking (AT) skills has become increasingly pronounced. Algorithmic thinking refers to the ability to analyse a problem and develop a step-by-step process to solve it.

This research investigates the efficacy of the Cross Array Task (CAT) as an assessment tool for AT skills within Switzerland's compulsory education system. Originally conceptualised as an *unplugged* activity, where students performed the task without digital technologies (e.g., by using gestures on paper) and an administrator manually assessed them, the CAT evolved into a *digital* activity that runs on an iPad. The CAT's digital transformation has automated the scoring of student responses and data collection, streamlining the assessment processes and facilitating efficient large-scale assessments. It has also enhanced scalability, making the CAT suitable for widespread use in educational settings. Furthermore, it provides immediate feedback to students and educators, supporting timely interventions and personalised learning experiences.

Our study aims to comprehensively investigate algorithmic competencies in compulsory education, examining their variations and influencing factors. This research examines key variables, such as age, sex, educational environment and school characteristics (e.g., the level and grade of education), and regional factors (e.g., the canton of the school) in Switzerland, and characteristics related to the specific assessment tool, including the type of artefact used, the complexity of the algorithms generated, and the level of autonomy. Additionally, it seeks to analyse the effectiveness of the unplugged and digital approaches in assessing AT skills, specifically comparing the unplugged and virtual CAT versions, aiming to provide insights into their advantages and potential synergies.

This investigation delineates the developmental progression of AT skills across compulsory education, emphasising the influence of age on algorithm development and problem-solving strategies. Furthermore, we reveal the impact of artefacts and the potential of digital tools to facilitate advanced AT skill development across diverse age groups. Finally, our investigation delves into the influence of school environments and sex disparities on AT performance, alongside the significant individual variability influenced by personal abilities and external circumstances.

These findings underscore the importance of tailored educational interventions and equitable practices to accommodate diverse learning profiles and optimise student outcomes in AT across educational settings.

* Corresponding author.
*E-mail addresses:* giorgia.adorni@idsia.ch (G. Adorni), igor.artico@usi.ch (I. Artico), alberto.piatti@supsi.ch (A. Piatti), elia.lutz@fhnw.ch (E. Lutz), luca.gambardella@idsia.ch (L.M. Gambardella), lucio.negrini@supsi.ch (L. Negrini), francesco.mondada@epfl.ch (F. Mondada), dorit.assaf@fhnw.ch (D. Assaf).

https://doi.org/10.1016/j.chbr.2024.100466
Received 10 June 2024; Received in revised form 31 July 2024; Accepted 31 July 2024
Available online 27 August 2024

## 1. Introduction

Computational thinking (CT) has become an essential skill for students in the 21st century, leading to increased efforts to integrate computer science (CS) education into K-12 classrooms (Weintrop, Rutstein, Bienkowski, & McGee, 2021). This initiative started with Jeannette Wing's introduction of the term CT, which she described as a fundamental skill for everyone, not just computer scientists (Wing, 2006). Wing emphasised that CT involves problem-solving processes that draw on concepts fundamental to computer science, such as abstraction, decomposition, and algorithmic design.

Despite the growth in tools, activities, and curricula for teaching CT, mainly due to the lack of a clear, universally accepted definition of CT, which poses challenges for its integration into educational standards and curricula (Piatti et al., 2022; Weintrop et al., 2021). For the purposes of this article, we refer to Piatti's definition of CT (Adorni et al., 2024d; Piatti et al., 2022): "CT is the cognitive activity required to solve problems through algorithms. CT involves three iterative steps: (1) setting a contextualised problem so that its solution can be computed (problem setting), (2) creating and representing an algorithm to be implemented by an agent (human, artificial, and/or virtual) to solve the problem (algorithm), and (3) assessing the solution's quality relative to the original problem (assessment)".

Within this broader context, there is a growing focus on algorithmic thinking (AT), which is recognised as a crucial component of CT in education (Wing, 2006, 2014, 2017; Yadav, Mayfield, Zhou, Hambrusch, & Korb, 2014). AT explicitly focus on the ability to design and express solutions to problems in a step-by-step, systematic manner (Adorni et al., 2024d). This focus on AT reflects its importance in developing logical reasoning, problem-solving skills, and the ability to structure and automate processes, which are essential in both computer science and other fields.

As AT becomes an essential part of compulsory education, systematically evaluating students' proficiency becomes necessary. This assessment is crucial for educators to understand how well their teaching methods and curriculum work in imparting AT skills. Secondly, it helps track individual student progress and eventually provides personalised feedback that tailors instruction to meet each student's diverse needs.

Despite its importance, assessing AT presents inherent challenges due to the lack of standardised tools and the diversity of evaluation methods available. Researchers have highlighted various approaches, including problem-solving activities, programming assignments, question tests, and scale surveys, each with its considerations and limitations (Adorni et al., 2024d; Ezeamuzie & Leung, 2021; Grover, 2017; Pilotti, Nazeeruddin, Mohammad, Daqqa, Abdelsalam, & Abdullah, 2022; Scherer, Siddiq, & Viveros, 2019). Establishing a universally accepted baseline for AT assessment remains challenging.

Our study is set within the context of Switzerland, a country with a unique educational landscape characterised by its multilingual environment, encompassing four national languages: German, French, Italian, and Romansh. This multilingualism influences various aspects of education, including developing and implementing assessment tools. Given this diversity, it is essential to design AT assessments that are accessible and effective across different linguistic regions. Furthermore, our focus spans across the spectrum of compulsory education, encompassing students from 3 to 16 years old. This broad age range underscores the need for versatile assessment tools that accommodate diverse developmental stages and educational contexts. In addressing this multifaceted challenge, our study employs the Cross Array Task (CAT) (Piatti et al., 2022), developed with multilingual capabilities, which assesses AT in both unplugged (non-digital) and digital environments, providing a comprehensive evaluation of students' skills and insights into the effectiveness of different instructional strategies.

To guide our investigation, we have formulated the following research questions:

RQ1. What are the baseline competencies in AT in compulsory education, and how do they develop across school grades?

RQ2. How do characteristics specific to the assessment instrument, such as different interaction modalities used in unplugged and digital instructional strategies, influence the development of AT skills in relation to sex, educational environment (e.g., school level and grade), and regional factors (e.g., the canton of the school)?

By addressing these specific research questions, our study aims to provide nuanced insights into AT skills development within compulsory education. While methodological considerations such as the need for robust and universal assessment methods have been addressed in previous literature (Adorni & Piatti, 2024c; Piatti et al., 2022), our study aspires to provide practical guidance for educators, researchers, and policymakers, fostering advancements in pedagogical approaches tailored for AT assessment. Additionally, we seek to offer a comprehensive view of how students engage with and benefit from both the unplugged CAT and the virtual CAT (the digital version), thereby providing insights into the effectiveness of each method.

## 2. Theoretical background

In this section, we provide an overview of theoretical frameworks related to algorithmic thinking (AT) assessment, empirical studies and research findings, and a detailed description of the Cross Array Task (CAT) assessment tool in both formats.

### 2.1. Foundations of algorithmic thinking

Early developmental psychologists such as Piaget and Vygotsky laid foundational theories on cognitive development, emphasising the role of active learning and social interactions in constructing knowledge during the early stages of childhood development (Piaget, 1964; Piaget, Cook, et al., 1952; Piaget & Mussen, 1983; Vygotsky, 1978). Piaget's constructivist theory posits that children build knowledge through hands-on experiences and interactions with their environment, while Vygotsky's social constructivist theory adds that social interactions and cultural context significantly influence learning outcomes. Both theories support the idea that engaging students in problem-solving and critical thinking activities, such as those involved in AT, can significantly enhance cognitive development.

Modern research extends this framework, highlighting the importance of early experiences in STEM education, underscoring the role of AT in cultivating skills crucial for future learning and proficient problem-solving in today's society (Georgiou & Angeli, 2021; Hsu, Chang, & Hung, 2018; Jiang & Wong, 2022; Kanaki & Kalogiannakis, 2022; Nikolopoulou & Tsimperidis, 2023; Voronina, Sergeeva, & Utyumova, 2016). The underlying belief is that by introducing AT concepts early on, students not only become acquainted with technology but also develop critical thinking, logical reasoning, and analytical skills that have transferable applications across various domains (Bers, Strawhacker, & Sullivan, 2022; Bocconi et al., 2022; Webb et al., 2017; Weintrop et al., 2021).

### 2.2. Challenges in assessing algorithmic thinking

Assessing AT poses significant challenges due to the absence of standardised tools and the diversity of evaluation methods available (Adorni et al., 2024d; Ezeamuzie & Leung, 2021; Grover, 2017; Pilotti et al., 2022; Scherer et al., 2019).

Empirical research has explored the effectiveness and challenges of these various AT assessment methods. Traditional assessment methods such as multiple-choice and closed-ended questions are widely used for their efficiency in covering broad topics, the straightforward administration and grading, but they may oversimplify the assessment by focusing on rote memorisation rather than deeper problem-solving skills (Campbell-Barr, Lavelle, & Wickett, 2012; Csernoch, Biró, Máth,

& Abari, 2015; Oyelere, Agbo, & Sanusi, 2022; Simmering, Ou, & Bolsinova, 2019; Wickey da Silva Garcia, Ronaldo Bezerr. Oliveira, & da Costa Carvalho, 2022). Conversely, open-ended and problem-solving tasks offer a more nuanced evaluation of algorithmic thinking by assessing students' reasoning and creativity, albeit at the cost of increased grading complexity (Csernoch et al., 2015). Similarly, programming assignments and coding challenges provide practical applications of algorithmic skills but require intensive grading efforts to evaluate code quality (Sun, Ouyang, Li, & Zhu, 2021). Robotic activities engage students in real-world problem-solving, offering tangible and interactive ways to assess AT and providing immediate feedback through interaction with robots, but may be limited by access to equipment (Keith, Sullivan, & Pham, 2019; McCormick & Hall, 2022). The best solution appears to be automatic assessment systems, which can efficiently scale to accommodate large numbers of participants and provide immediate and consistent feedback, yet they are still developing in their ability to comprehensively assess complex AT skills, especially in monitoring learners' progress over time (Qian & Lehman, 2018; Romero, Lepage, & Lille, 2017; Stanja, Gritz, Krugel, Hoppe, & Dannemann, 2022).

### 2.3. Importance of comparing unplugged and digital approaches

The coexistence of unplugged and digital assessment methods further complicates AT assessment. Unplugged methods, characterised by their hands-on, non-digital nature, involve tangible activities to assess fundamental CS concepts in an accessible and engaging manner (Adorni et al., 2024d; Bell, Alexander, Freeman, & Grimley, 2009; Brackmann, Romá'n-González, Robles, Moreno-León, Casali, & Barone, 2017; Del Olmo-Muñoz, Cózar-Gutiérrez, & González-Calero, 2020; Piatti et al., 2022). These methods involve physical activities and exercises that teach CS principles without using computers. In contrast, digital methods utilise technological tools and software to engage students in programming tasks and other computer-based activities to assess AT skills within a digital context. These methods use software and digital tools to engage students in programming and other computer-based tasks.

Unplugged activities excel in building a strong foundation in computational principles, promoting clear concepts, and encouraging collaborative learning, but they may lack exposure to digital problem-solving and scalability for large-scale automatic assessments (Del Olmo-Muñoz et al., 2020; El-Hamamsy, Zapata-Cáceres, Barroso, Mondada, Zufferey, & Bruno, 2022; Piatti et al., 2022). On the other hand, while digital methods offer flexibility, interactive, and individualised learning experiences with immediate feedback, they may have limitations. Digital methods might lack the tangible, hands-on engagement that unplugged activities provide, potentially resulting in less accessible or engaging learning experiences for certain learners (Piatti et al., 2022; Relkin, de Ruiter, & Bers, 2020; Román-González, Pérez-González, & Jiménez-Fernández, 2017; Zapata-Cáceres, Martín-Barroso, & Román-González, 2020). Additionally, their reliance on digital platforms could introduce barriers for students with limited access to technology or those who prefer non-digital learning environments (Bell & Vahrenhold, 2018; Brackmann et al., 2017; Kalelioglu, Gulbahar, & Kukul, 2016; Relkin et al., 2020). However, the scalability of digital methods remains advantageous for automatically collecting data in large-scale educational assessments.

The comparison between unplugged and digital methods is crucial for educators and researchers to identify which strategies are most effective in different educational settings for assessing AT skills. This analysis can guide the establishment of standardised evaluation criteria, ensuring consistency and fairness in assessing students' computational competencies over time.

For this article, we adopt the Cross Array Task (CAT) as our assessment instrument due to its versatility in exploring AT skills across both non-digital and digital contexts, addressing the evolving needs of modern education. The CAT incorporates an automatic assessment system in its digital format, enabling scalable and consistent evaluation of students' AT abilities. If both unplugged and digital methods are found to be equally effective, it would validate the CAT as a reliable assessment tool, demonstrating its effectiveness in both contexts.

Understanding the strengths and limitations of both approaches enables educators to tailor their instructional methods to meet the diverse learning needs and preferences of students, thereby enhancing educational outcomes. Establishing a universally accepted benchmark for assessing AT skills is crucial for tracking progress and fostering continuous improvement in computer science education. A standardised approach facilitates comparisons across various educational contexts and supports evidence-based decision-making in curriculum development and educational policy.

### 2.4. The CAT assessment instrument

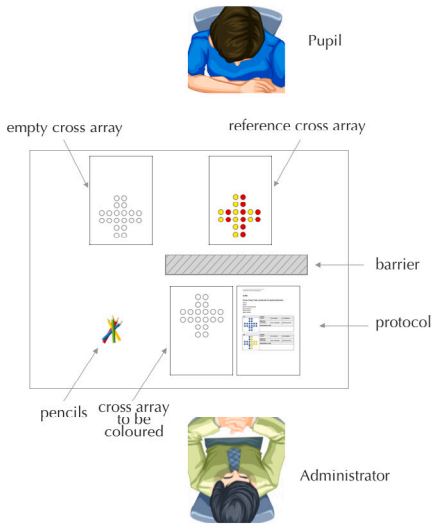
The CAT is an educational activity designed to evaluate algorithmic skills in students along the entire compulsory school (Piatti et al., 2022). In this assessment, each student is assigned 12 tasks, each requiring them to formulate a set of instructions known as an algorithm (See Fig. A.1). The objective is to describe and replicate unique cross arrays schema characterised by a pattern of 20 dots forming a 2-thick cross, with colours selected from a set of options (typically yellow, green, blue, or red) and featuring diverse regularities and patterns.

Fig. 1 depicts the activity experimental setting. Initially conceived as an unplugged activity, the CAT involved face-to-face interaction between the problem solver and a human agent. The pupil has at his disposal two types of cognitive artefacts to communicate algorithms: verbal instructions (V) or accompanying the voice with gestures on an empty sample schema (VS). At the same time, the human administrator manually interprets, replicates and records all pupil instructions. A physical barrier initially promoted pupil autonomy during this interaction by concealing the administrator's replication process. Its removal provides direct visual feedback for the pupil, thus reducing his autonomy level.

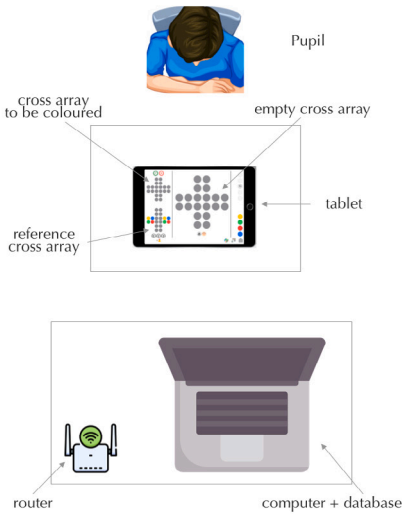
To streamline the assessment process, enabling large-scale assessment, address the time-consuming nature of individual administration, and reduce inconsistencies in evaluations due to human involvement, a digital version of this activity, called virtual CAT, was developed (Adorni & Piatti, 2024c; Adorni, Piatti, & Karpenko, 2024e). This transition replaced the human administrator with a virtual agent, transforming the interaction into a device-mediated process. The virtual agent interprets instructions using a programming language interpreter, automating algorithm recording and evaluation to provide immediate feedback and guidance. Administration via individual devices such as tablets enables simultaneous participation by multiple students, enhancing accessibility and efficiency.

Cognitive artefacts evolved in a digital version, called virtual CAT, replacing direct verbal (V) and gestural communication (VS) with a visual programming interface (P), where users create algorithms by arranging visual coding blocks via drag-and-drop (see Fig. A.2), and a gesture-based interface (G) where users manipulate and colour the cross directly by touching dots (see Fig. A.3). Both versions of the activity support varying levels of student autonomy through the option for students to request visual cues on their colouring progress. The digital format also introduces enhancements like task skipping or restarting, further optimising the task experience.

To accommodate Switzerland's multilingual context, the virtual CAT included support for multiple languages – Italian, French, and German – thus ensuring broader accessibility for users of different native languages and facilitating adoption across various educational institutions. An English version is also available to extend the application's utility to a broader audience.




**(a) Unplugged CAT.** A pupil verbally guides the administrator in recreating a coloured reference cross array possibly supported by gestures on an empty initially with a removable physical barrier preventing visual cues. The administrator interprets and protocols all pupil's instructions.




**(b) Virtual CAT.** A pupil uses a gesture-based or visual block-based programming interface to recreate a coloured reference cross array schema, initially without visual feedback due to a toggleable feature. The system automatically interprets and logs all actions and algorithms in an external database.

**Fig. 1.** Experimental setting of the CAT.
*Source:* Adapted from Adorni and Piatti (2024c), Piatti et al. (2022).

#### 2.4.1. Algorithmic skills assessment

The CAT assessment instrument evaluates students' algorithmic skills through a structured approach that assesses both the complexity of algorithms generated and their efficiency in design. Here, we provide a detailed exploration of how the CAT measures algorithmic proficiency, from basic operations to more advanced patterns, and introduce the adjusted algorithm dimension metric to enhance evaluation accuracy.

*Algorithm dimension.* The algorithm dimension serves as an indicator of the complexity of algorithms generated by students during the CAT assessment. Each operation within the algorithm is evaluated based on its level of complexity, which ranges across three distinct levels. These levels signify increasing degrees of algorithmic sophistication, with higher scores indicating greater proficiency and the ability to handle more intricate tasks effectively.

0D: this level entails colouring individual dots within the cross;

1D: this level entails colouring multiple dots with the same colour, following patterns such as rows or diagonals;

2D: this level entails creating more complex patterns involving alternating colours, repetitions of patterns, and other intricate arrangements.

The score range from 0 for the first level (0D) to 2 for the most complex level (2D). The overall algorithm dimension is determined by the most complex operation successfully performed by the student during the assessment. A higher algorithm dimension not only indicates greater proficiency but also implies the ability to handle both complex and simpler operations effectively, reflecting the student's depth of understanding in algorithmic thinking and problem-solving

*Adjusted algorithm dimension.* Acknowledging the need to assess algorithm efficiency alongside complexity, we have introduced an adapted metric that considers the number of commands used, providing a more nuanced evaluation of students' algorithmic competencies. It recognises cases where a simpler yet more efficient algorithm may perform better than a complex one with more commands. The adjusted score, denoted as $\widetilde{\text{AD}}$, is calculated using a formula that balances the highest complexity level achieved by the student against the overall workload:

$$\widetilde{\text{AD}} = \frac{1 + P_{\text{max-}d} + \sum_d \left(C_d \cdot P_d\right)}{C_{\text{total}}}, \tag{1}$$

where, $d$ is the complexity level of the algorithm (i.e., 0, 1, or 2); $P_{\text{max-}d}$ are the points assigned to the highest complexity level used by the student, computed as the original algorithm dimension score plus one; $C_d$ is the number of commands at complexity level $d$; $P_d$ are the points for the complexity level $d$, computed as the original algorithm dimension score at that complexity level plus one; $C_{\text{total}}$ is the overall number of commands used across all levels. The first term in the formula (1) gives a score for the most complex algorithm achieved by the student, adjusted for the total commands used, favouring higher-level algorithms but considering the overall workload in terms of the number of commands executed. The second term calculates a weighted score for each complexity level, factoring in the proportion of commands used at each complexity level relative to the total command count and multiplying it by the points for that level.

#### 2.4.2. Interaction dynamics assessment

The interaction dimension within the CAT assessment evaluates how students engage with the assessment instrument, reflecting both the complexity of the artefacts used and the level of autonomy demonstrated during task execution, determined by the extent to which they asked for visual cues and relied on visual feedback.

This dimension considers various modalities of interaction, whether unplugged or virtual, each presenting distinct levels of complexity and autonomy.

*Interaction dimension (for the unplugged CAT).* For activities conducted without digital interfaces, the interaction dimension includes three levels of complexity:

VSF: this level involves using voice commands and hand gestures on an empty cross array, hinging on visual feedback;

VS: this level involves using voice commands and hand gestures on an empty cross array, without hinging on visual feedback;

V: this level involves using only voice commands without hand gestures on an empty cross array or visual feedback;

Scores range from 0 for the simplest level of complexity (VSF) to 2 for the highest level (V), based on the complexity of interaction observed during the assessment.

*Interaction dimension (for the virtual CAT).* When tasks are conducted through digital interfaces, the interaction dimension expands to four levels of complexity:

GF: this level involves using the gesture interface, hinging on visual feedback;
G: this level involves using the gesture interface, without hinging on visual feedback;
PF: this level involves using the visual programming interface, hinging on visual feedback;
P: this level involves using the visual programming interface, without hinging on visual feedback;

Scores range from 0 for the simplest level of complexity (GF) to 3 for the highest level (P), based on the complexity of interaction observed during virtual assessments.

The interaction dimension complements the algorithm dimension by providing a comprehensive view of students' engagement preferences and capabilities across different interaction modalities. While the algorithm dimension focuses on the highest complexity level achieved, for the interaction dimension we report the lowest complexity level reached and the predominant interaction style used throughout the assessment. This dual approach offers valuable insights into students' interaction patterns, autonomy, adaptability, and proficiency in utilising various interaction methods during the assessment tasks. A higher interaction dimension indicates a greater level of complexity in the student's interaction with the activity, suggesting that the student used more complex artefacts or demonstrated a higher level of autonomy during the task.

### 2.4.3. CAT competencies assessment

The CAT assessment evaluates students' skills by measuring both algorithmic complexity and interaction dynamics, combining these aspects into a single CAT score metric to provide a comprehensive assessment of their performance. Table 1 illustrates how this metric is computed, enabling us to assess task performance across both unplugged and digital approaches. While algorithm dimensions are directly comparable since they share the same levels between the two approaches, interaction dimensions are not. Nonetheless, they provide insights into the progression from lower to higher levels of interaction complexity.

**Table 1**
**CAT score metric to assess task performance.** Rows represent algorithm dimensions, while columns indicate interaction dimensions.

**(a) Unplugged CAT.** The interaction dimensions correspond to the use of voice and hand gestures on an empty cross array, hinging on visual feedback (VSF), the use of voice and hand gestures on an empty cross array without visual feedback (VS), and the use of voice alone without hand gestures or visual feedback (V).

| | VSF | VS | V |
|---|---|---|---|
| **0D** | 0 | 1 | 2 |
| **1D** | 1 | 2 | 3 |
| **2D** | 2 | 3 | 4 |

**(b) Virtual CAT.** The interaction dimensions correspond to the use of gesture interface hinging on visual feedback (GF), the use of gesture interface without hinging on visual feedback (G), the use of visual programming interface hinging on visual feedback (PF), and the use of visual programming interface without hinging on visual feedback (P).

| | GF | G | PF | P |
|---|---|---|---|---|
| **0D** | 0 | 1 | 2 | 3 |
| **1D** | 1 | 2 | 3 | 4 |
| **2D** | 2 | 3 | 4 | 5 |

### 2.4.4. Task metrics

The CAT assessment instrument evaluates also various metrics to gauge students' proficiency in AT and task execution. Key metrics assessed include:

*Participation rate.* The participation rate measures whether students attempted and concluded each task assigned during the CAT assessment, regardless of correctness. Each student is assigned 12 tasks, and the participation rate indicates how many of these tasks. This metric provides an initial overview of students' engagement and persistence in the assessment activities.

*Success rate.* The success rate evaluates the number of tasks that students correctly solved during the CAT assessment, irrespective of efficiency or the number of attempts made.

*Number of restarts.* The number of restarts reflects students' approach to problem-solving, particularly their use of trial and error (T&E) strategies. It counts instances where students choose to restart the tasks, indicating their iterative approach to refining algorithms and achieving desired outcomes.

*Efficiency.* Efficiency evaluates how effectively students complete tasks, considering the time taken as factor.

### 2.4.5. Validity and reliability

The CAT was validated as an unplugged task in spring 2022 (Piatti et al., 2022) and further tested in a digital format as virtual task during a subsequent pilot phase in spring 2023 (Adorni & Piatti, 2024c).

Ensuring the validity and reliability of the CAT involved several steps. Initially, the CAT was designed and assessed within the CT-cube framework (Piatti et al., 2022), which extends traditional CT models to include developmental and situational aspects of skills. The design process for both versions prioritised usability by establishing clear objectives and analysing requirements. Especially for the digital version of the activity, prototypes were meticulously developed, focusing on user experience and undergoing thorough testing, including expert UX inspections. Children actively participated in the design process, serving as both informants and evaluators. Their input ensured that the CAT aligned with their needs, guaranteeing functionality, accessibility, and engagement.

Practical test in real educational settings assessed both versions of the CAT across diverse K-12 student populations, ensuring the activity accommodated various age groups and backgrounds effectively. These tests pinpointed areas for improvement and validated the platform's readiness for broader implementation. By observing pupils' proficiency in generating algorithms across various tasks and settings, evaluating their ability to perform the task using all available artefacts, regardless of age or background, by assessing their consistency in completing provided schemas and measuring engagement and success rates, it was confirmed that the instrument consistently elicited active participation and yielded reliable outcomes. Observations included how students approached and completed tasks, their diverse algorithmic strategies, and the tool's ability to yield consistent outcomes under different conditions.

Factors affecting reliability, such as the length of the assessment, the suitability of the tasks for the students being assessed, and the consistency in test administration, were all considered and addressed in both studies. Insights gained from these assessments were essential in refining the CAT to better serve the varied needs of students and educators.

These steps collectively affirmed the platform's capability to deliver valid and reliable assessments in educational contexts. Consequently, the CAT can be recognised as a dependable tool for assessing AT skills. Its consistent outcomes inspire confidence in making broad assertions about students' achievement levels across multiple assessments.

## 3. Methods

In this section, we outline the methodologies used to assess students' algorithmic skills using the Cross Array Task (CAT) in both unplugged and virtual formats. Our overview encompasses experimental settings, data collection procedures, participant selection, and data analysis methodologies.

**Table 2**
**Swiss compulsory education structure under the HarmoS Agreement.** Detailed representation of the Swiss compulsory education system, correlating HarmoS grades (HGs) and ages with stages in German-speaking (DE), French-speaking (FR), and Italian-speaking (IT) cantons. The layout showcases three key educational cycles: preschool (Kindergarten/Cycle primaire 1/Scuola dell'infanzia), primary school (Primarschule/Cycle primaire 2/Scuola elementare), and lower secondary school (Sekundarstufe I/Cycle secondaire/Scuola media), each mapped to specific grades and regions.

| **Harmos grade** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age | 3-4 | 4-5 | 5-6 | 6-7 | 7-8 | 8-9 | 9-10 | 10-11 | 11-12 | 12-13 | 13-14 | 14-15 |
| DE | - | Kindergarten | | Primarschule | | | | | | Sekundarstufe I | | |
| FR | - | Cycle 1 (primaire) | | | | Cycle 2 (primaire) | | | | Cycle 3 (secondaire 1) | | |
| IT | Scuola dell'infanzia | | | Scuola elementare | | | | | Scuola media | | | |

### 3.1. Experimental settings

For both the unplugged and virtual CAT studies, there were distinct logistical arrangements and roles of administrators in guiding and assisting students during the activities.

For the unplugged CAT, two pupils at a time were randomly selected from the class and taken to a separate room to minimise interference with the remaining students. The administrator provided initial instructions and assistance during the activity, such as suggesting simpler methods when students encountered difficulties, like complementing voice instructions with gestures or relying on visual feedback. The time required to solve all the 12 schemas varied from a minimum of 10 min – in the case of the older pupils – to a maximum time of 45 min – for the younger ones. Approximately, the total time required to administrate the CAT to all 109 has been 36 h.

For the virtual CAT, a training module was integrated into the app to familiarise students with the assessment tool. Training sessions, lasting about 30–45 min, were conducted in groups based on device availability. Each student was provided with individual devices, allowing the activity to be orchestrated for the entire class simultaneously and seamlessly. During the actual validation phase, the administrator refrained from providing help or tutoring.

### 3.2. Data collection

In the data collection process, we recorded various factors related to the student, the educational environment, and the session context that may influence students' performance and competency development in the context of assessing students' algorithmic skills using the CAT. These factors were considered across both unplugged and virtual formats of the activity, referred to as domains. Understanding them is crucial for interpreting the data and evaluating performance comprehensively.

#### 3.2.1. Factors affecting performance

*Domain.* We differentiate between the unplugged and digital formats of the activity, recognising that each format involves different interaction methods which can affect student performance and engagement.

*Canton.* Given Switzerland's linguistic diversity, we categorise students based on their respective geographic origin or the canton where they live (similar to federal states) to explore potential variations in AT influenced by each unique linguistic region's educational practices.

*Educational context.* To understand the educational context, we examine the combination of two key factors: sessions and HarmoS grades (HGs).

Sessions refer to distinct instances or implementations of the educational activity, each characterised by unique student compositions, class dynamics, and educational settings.

HGs, on the other hand, denote specific levels within Switzerland's federalist education system, In particular, the HarmoS Agreement provides a standardised yet flexible structure for learning across linguistic regions (Swiss Conference of Cantonal Ministers of Education, 2007; UNESCO Institute for Statistics, 2012). Table 2 provides an overview of the Swiss compulsory education system, spanning eleven years (twelve in Ticino), including preschool, primary, and lower secondary levels, with variations tailored to the linguistic regions of Switzerland, German, French, and Italian-speaking parts.

*Specific school institutions.* We account for the influence of specific school institutions, aiming to understand variations in AT performance across different schools and recognise the impact of institution-specific dynamics.

*Age category.* We categorised students into four age groups, 3–6, 7–9, 10–13, and 14–16 years old, to understand how algorithmic capabilities evolve across their entire compulsory education journey, similar to our approach in the unplugged CAT experimentation (Piatti et al., 2022).

*Sex.* We delve into the influence of sex on students' performance in AT, recognising potential disparities and their implications.

#### 3.2.2. Data collection process

During the data collection process, the administrator manually recorded session and participant details in both studies. The same session information was collected for contextualisation and analysis purposes. Each session was assigned a unique identifier, and specific details, such as the date, canton, school name and type, and the students' HarmoS grade (HG) level, were recorded. Student information was limited to sex and date of birth, with birth dates used to calculate ages, a significant factor in our demographic analysis. To protect student privacy, unique identifiers were assigned to each participant, keeping the data anonymous and secure.

In the unplugged CAT study, administrators manually recorded activity and performance details, including operations performed, algorithm complexity, and artefact type.

In the virtual CAT study, the application automatically tracked activity information. It logged every action performed by the students, including a complete record of the types of operations executed. These actions ranged from basic tasks such as adding, confirming, or removing commands to more complex actions like updating command properties, resetting algorithms, and altering modes of interaction. The application also tracked instances of tasks being completed or abandoned, marking each with a timestamp and noting the specific nature of the operation. This extensive logging process provided an in-depth view of student engagement and interaction patterns with the application.

All data collected in this study were pseudonymised, aligning with prevailing open science practices in Switzerland (SNSF, 2021). This step protected participants' privacy while allowing the data to be accessible and analysed within the academic community. Data from the unplugged study is not available online due to some constraints specified in the original consent form but is securely stored on protected servers and accessible only to researchers directly involved in the project. Data from the virtual study has been made publicly accessible on the Zenodo platform, with identifiable information like school names and class details omitted to safeguard participant confidentiality while enabling access to other scholars for further studies (Adorni, 2024a).

**Table 3**
**Demographic analysis of students in the unplugged and digital studies.** Overview of student demographics by session, including canton, school ID and type, HarmoS grade (HG), age category (mean age and standard deviation), and sex distribution (number of female and male students). Note that while Preschool A took part in both studies, the pupils involved were not the same.

**(a) Unplugged CAT.**

| Session | Canton | School ID & type | HG | Age category | Female | Male | Total |
|---|---|---|---|---|---|---|---|
| 1U | Ticino | A Preschool | 0, 1, 2 | 3-6 yrs ($\mu$ 4.9 ± 0.9 yrs) | 8 | 13 | 21 |
| 2U | Ticino | B Primary school | 3 | 3-6 yrs ($\mu$ 6.7 ± 0.5 yrs) | 4 | 8 | 12 |
| 3U | Ticino | B Primary school | 5 | 7-9 yrs ($\mu$ 8.7 ± 0.6 yrs) | 7 | 8 | 15 |
| 4U | Ticino | B Primary school | 7 | 10-13 yrs ($\mu$ 10.5 ± 0.6 yrs) | 8 | 11 | 19 |
| 5U | Ticino | C Lower secondary school | 9 | 10-13 yrs ($\mu$ 12.5 ± 0.5 yrs) | 8 | 7 | 15 |
| 6U | Ticino | C Lower secondary school | 10 | 14-16 yrs ($\mu$ 13.0 ± 0.0 yrs) | 5 | 2 | 7 |
| 7U | Ticino | C Lower secondary school | 11 | 14-16 yrs ($\mu$ 14.5 ± 0.7 yrs) | 9 | 5 | 14 |
| 8U | Ticino | C Lower secondary school | 11 | 14-16 yrs ($\mu$ 14.5 ± 0.5 yrs) | 2 | 4 | 6 |
| | | | | ($\mu$ 9.9 ± 3.5 yrs) | 51 | 58 | 109 |

**(b) Virtual CAT.**

| Session | Canton | School ID & type | HG | Age category | Female | Male | Total |
|---|---|---|---|---|---|---|---|
| 1V | Ticino | A Preschool | 0, 1, 2 | 3-6 yrs ($\mu$ 5.0 ± 0.8 yrs) | 6 | 7 | 13 |
| 2V | Solothurn | D Preschool | 2 | 3-6 yrs ($\mu$ 5.9 ± 0.3 yrs) | 8 | 6 | 14 |
| 3V | Ticino | E Primary school | 4 | 7-9 yrs ($\mu$ 7.7 ± 0.6 yrs) | 7 | 8 | 15 |
| 4V | Solothurn | D Primary school | 6 | 7-9 yrs ($\mu$ 9.9 ± 0.3 yrs) | 8 | 10 | 18 |
| 5V | Ticino | F Lower secondary school | 8 | 10-13 yrs ($\mu$ 11.6 ± 0.5 yrs) | 11 | 9 | 20 |
| 6V | Ticino | F Lower secondary school | 10 | 14-16 yrs ($\mu$ 13.9 ± 0.8 yrs) | 8 | 5 | 13 |
| 7V | Ticino | G Lower secondary school | 10 | 14-16 yrs ($\mu$ 13.6 ± 0.6 yrs) | 7 | 7 | 14 |
| 8V | Ticino | G Lower secondary school | 11 | 14-16 yrs ($\mu$ 14.7 ± 0.5 yrs) | 6 | 5 | 11 |
| 9V | Solothurn | D Lower secondary school | 11 | 14-16 yrs ($\mu$ 15.5 ± 0.5 yrs) | 4 | 7 | 11 |
| | | | | ($\mu$ 10.7 ± 3.6 yrs) | 65 | 64 | 129 |

### 3.3. Participants selection

This study examines two distinct participant groups, each associated with either the unplugged or digital approach. The first group, drawn from the experimental study conducted between March and April 2021 to assess the unplugged CAT (Piatti et al., 2022), consisted of 109 students (51 girls and 58 boys) sampled from eight classes in three public schools in Ticino. The second group, participating in the experimental study conducted in Spring 2023 to assess the virtual CAT, comprised a more extensive and diverse sample of 129 students (65 girls and 64 boys), selected from nine classes across five public schools in Ticino and Solothurn cantons.

It is important to note that the school and class selection process was not random; rather, they were contacted and agreed to take part in the study. This approach aimed to include a broader demographic of students, ensuring diversity in age, sex, and geographic origin across both linguistic regions.

Table 3 provides a breakdown of participant attributes, organised by session (e.g., 1U for the unplugged study, 1 V for the virtual one), including factors like canton, school ID and type, HarmoS grade (HG), age category (with mean age and standard deviation), and sex distribution. While a balanced distribution is evident across various factors like school type, HG, age category, and sex, there is a notable exception with the canton representation. Specifically, there are fewer students from Solothurn Canton compared to Ticino, indicating a slight imbalance in geographic representation. Nonetheless, these demographic analyses provide valuable insights into the diverse characteristics of participants in both studies.

This study adhered to high ethical standards, especially considering the involvement of young participants. We prioritised transparency and respect in all procedures involving pupils, parents, and educational institutions (Aebi-Müller, Blatter, Brigger, Constable, Eglin, Hoffmeyer, Lautenschütz, Lienhard, Pirinoli, Röthlisberger, & Spycher, 2021; Petousi & Sifaki, 2020). Initially, comprehensive documentation outlining the study's objectives, data collection and storage methods, and details about the research team were provided to school directors, teachers, and parents. Explicit authorisation was obtained from school directors and teachers for the study activities. Subsequently, informed consent was diligently secured from parents, explicitly seeking permission for their children's participation and data use.

### 3.4. Data analysis

To address our research question regarding the evaluation of AT competency in learners and the factors affecting performance, we employed a variety of statistical techniques. These methods were chosen based on their appropriateness for analysing the CAT's efficacy as an assessment tool in both unplugged and virtual formats. Python was used for exploratory and descriptive analyses, while R facilitated advanced statistical analyses (R Core Team, 2023; Van Rossum & Drake, 2009).

#### 3.4.1. Algorithmic thinking skills development

To analyse how students' interaction strategies influence their algorithmic thinking skills, we examined data from both unplugged and virtual CAT experiments, focusing on the choice and use of artefacts, levels of autonomy, and their effects on algorithmic thinking.

Initially, we conducted a qualitative analysis of students' interaction strategies to uncover recurring patterns and strategies across different age groups. This involved examining how students engaged with various artefacts and how these interactions related to their algorithmic thinking skills. Descriptive statistics were computed to summarise the distribution of algorithmic dimensions across interaction types and age groups, providing a baseline for understanding general trends in algorithmic skills development.

To validate and extend our qualitative findings, we performed a series of statistical analyses. We began with an Analysis of Variance (ANOVA) to assess whether the type of artefact used significantly affected algorithmic complexity. This method lets us compare the average algorithmic complexity across multiple artefact types to see if there are any significant differences. This analysis was conducted separately for both unplugged and virtual environments. To find specific differences in complexity between students using different artefacts, we used t-tests (Bartlett, 1937; Chambers, Hastie, & Pregibon, 1990; Cox & Hinkley, 1979; Davison & Hinkley, 1997; Hastie, Friedman, & Tibshirani, 2001; James, Witten, Hastie, & Tibshirani, 2013; Silvey, 2017). This test compares the means of two groups to see if they differ significantly, helping us identify which artefacts were most effective. After finding significant differences with ANOVA, we conducted post-hoc analyses to pinpoint specific differences between artefacts and highlight

which were most effective in fostering complex algorithms. Tukey's HSD tests with Benjamini–Hochberg adjustments helped control for errors when comparing multiple groups, airwise t-tests with Bonferroni correction were used to ensure accurate comparisons between individual artefacts, while chi-squared tests of proportions assessed categorical data (Benjamini & Hochberg, 1995; Bland & Altman, 1995; Cochran, 1954; Dunn, 1961; Moore & McCabe, 1989; Newcombe, 1998a, 1998b; Perneger, 1998; Sedgwick, 2014; Tukey, 1949; Wilson, 1927; Yates, 1934).

For the virtual CAT interaction strategies, we tracked the frequency of different interaction dimensions to identify preferences and tendencies in artefact usage and autonomy among various student groups. This analysis revealed how students adapt and switch between modalities, providing insights into their capabilities, preferences, and adaptability.

To explore how interaction strategies and demographic factors jointly impact algorithmic skills, we examined interaction effects between artefacts, age, sex, and schemas. To understand age-specific patterns in algorithmic complexity we used ANOVA, which helped us compare how age groups and interaction types affected algorithm complexity, and linear regression models, which showed how these factors combined to influence algorithmic skills (Fisk & Weisberg, 1982; Hastie et al., 2001; Martin & Maes, 1979; Seber, 1984).

We extended our investigation with Estimated Marginal Means (EMMs) analysis to examine the combined effect of age category and interaction dimension on algorithm complexity in both unplugged and virtual environments. (Lenth, 2023). This method helped interpret the interaction between age and the influence of interaction dimensions on algorithm complexity. We also used chi-squared tests of proportions to explore differences in the distribution of higher algorithm dimensions (2D) across various age categories and interaction dimensions.

Finally, to understand the interplay between age and sex, we analysed how these factors interact and affect algorithmic skills using ANOVA and linear regression models. These models were consistent with those used in previous analyses, ensuring a comprehensive understanding of the impact of age and sex on AT.

#### 3.4.2. Student participation and success

We use descriptive statistics to explore participation and success rates across various schemas, age groups, and interaction dimensions, providing an overview of the overall trends and distributions of student performance within each category.

To assess the significance of performance differences between age categories and domains (unplugged and digital), we employed several statistical tests. Chi-square test was used to evaluate the relationship between categorical variables, such as age categories and interaction dimensions. This test allowed us to determine if the observed frequencies of student success rates differed significantly from what we would expect by chance, thereby revealing any notable differences in participation and success across different groups.

Tukey's HSD test was used to identify which specific age groups differed in their success rates, across domains. To address the issue of multiple comparisons and reduce the likelihood of Type I errors (false positives), we applied the Benjamini–Hochberg correction. This adjustment modifies the p-values to control the false discovery rate, ensuring that our findings are reliable and accurate.

#### 3.4.3. Trial and error strategies

We investigated students' trial and error (T&E) behaviours within the virtual domain, focusing on understanding the factors influencing task restarts and their correlation with task success. We employed Ordinary Least Squares (OLS) regression (Hastie et al., 2001; James et al., 2013; Stone & Brooks, 1990; Zdaniuk, 2014), to explore the relationship between T&E behaviours, specifically the frequency restarts, and several predictors, including task complexity, age, sex, and interaction dimension. This method was chosen because it allows us to assess how well these predictors explain variations in restart frequency, quantifying the strength and direction of these relationships, and determine which factors most significantly impact students' tendency to restart tasks and how these factors correlate with overall task success.

#### 3.4.4. Linear mixed model assessment of student performance

To evaluate the influence of various factors on CAT scores, we employed linear mixed models (LMMs) due to the hierarchical nature of our data, where students are nested within sessions, schools, and cantons (Hox, Moerbeek, & Van de Schoot, 2017; Raudenbush & Bryk, 2002). LMMs were chosen because they can handle such nested structures and account for both fixed effects (e.g., sex) and random effects (e.g., variability across sessions, schools, and cantons). This approach allowed us to understand how different predictors influence CAT performance while appropriately addressing the data's hierarchical correlations.

We build a baseline model (M0) using the Restricted Maximum Likelihood (REML) approach with Satterthwaite's approximation for degrees of freedom, implemented via the `lmer` function from the `lmerTest` packages in R (Kuznetsova, Brockhoff, & Christensen, 2017).

$$\begin{aligned}\texttt{CAT\_SCORE} = &\beta_0 + \beta_1 \cdot \texttt{CANTON} + \beta_2 \cdot \texttt{SEX} + \\ &+ u_{\texttt{STUDENT}} + u_{\texttt{SESSION\_GRADE}} + u_{\texttt{SCHOOL}} + u_{\texttt{SCHEMA}} + \epsilon.\end{aligned} \tag{2}$$

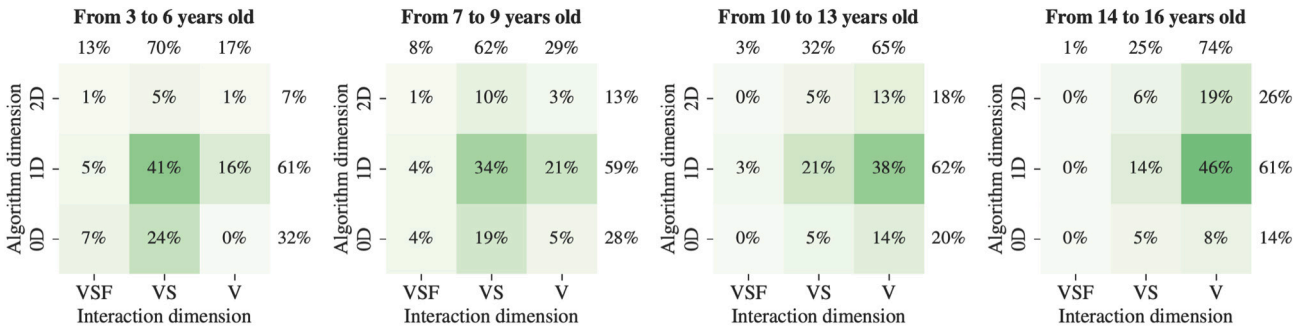
In our model, defined in (2), we considered various components. The outcome variable is the `CAT_SCORE`, representing the performance score. As fixed effects, we included `CANTON` due to the limited number of cantons (only two cantons) and `SEX`, a binary predictor for sex. As random effects, we included `STUDENT`, accounting for natural heterogeneity among students, acknowledging unique factors like abilities, prior knowledge, and unobserved characteristics inherent to each student; `SESSION_GRADE` captures variations related to HarmoS Grade and testing sessions, such as time of day, classroom conditions, and differences introduced by various teachers, as well as disparities across educational levels, acknowledging each grade's unique curricular and teaching aspects; `SCHOOL` represents variability among different schools, encompassing their unique environments and resources; and `SCHEMA` accounts for variability among the 12 distinct tasks, isolating the task-specific characteristics. $\beta_0$, $\beta_1$, and $\beta_2$ are the coefficients for the fixed effects, while $\epsilon$ is the error term representing the model's unexplained variability.

The data analysis methods employed include iterative model comparison and refinement, starting with the baseline model (M0). We used the Likelihood Ratio Test (LRT) via the `anova` method in the `lmreTest` package in R (Bartlett, 1937; Chambers et al., 1990; Kuznetsova et al., 2017) to assess different model fits and determine predictors' contributions to student performance. The LRT allowed us to compare nested models to identify which predictors significantly improved the model fit.

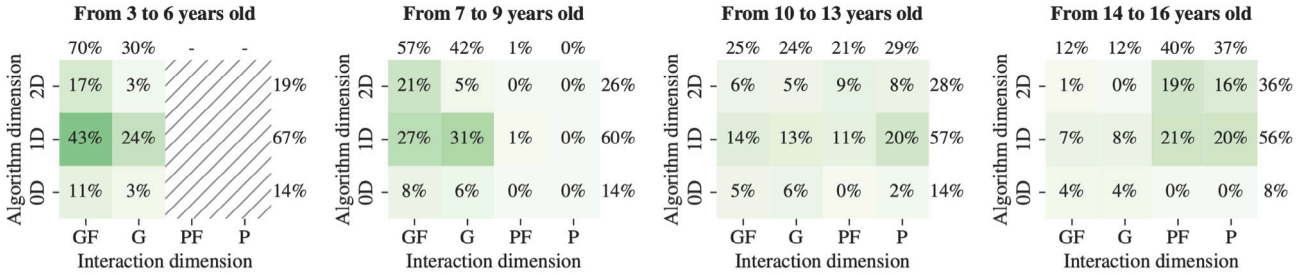
After selecting the final model, we conducted a type III Analysis of Variance (ANOVA) with Satterthwaite's method. This analysis assessed the significance of different terms in explaining the variability in CAT scores across both virtual and unplugged domains. Satterthwaite's method is used to adjust for unequal variances and sample sizes, providing a more accurate test of significance for the predictors.

In addition, we investigated the effect of task completion time on performance, quantitatively examining its relationship with variables like age and interaction types. We analysed completion time as a random effect to explore its correlation with performance. Treating it as a random effect allowed us to better capture individual variability in completion pace, task complexity, and concentration levels, which might otherwise skew results if treated as a fixed effect.

We also explored student performance dynamics in both virtual and unplugged settings by merging the data and introducing new fixed-effect predictors to differentiate between the two domains. Furthermore, we conducted LRT to assess the overall impact of sex, isolating it from the variability linked to different schools. This helped us understand how sex influences performance while accounting for school-related differences.




**(a) Unplugged CAT.** The interaction dimensions correspond to voice with hand gestures on an empty cross array, hinging on visual feedback (VSF), voice with hand gestures on an empty cross array (VS), and voice alone (V)




**(b) Virtual CAT.** The interaction dimensions correspond to gesture interface hinging on visual feedback (GF), gesture interface (G), visual programming interface hinging on visual feedback (PF), and visual programming interface (P). It's worth noticing that the younger age category was not allowed to use the visual programming interfaces (PF and P), so the only reported interaction dimensions are GF and G.

**Fig. 2. Interaction strategies and algorithmic skills development.** The *y*-axis illustrates the distribution of algorithmic dimensions across interaction dimensions on the *x*-axis for each age category. Percentages represent the proportion of each combination within their respective age groups. The sum of percentages across rows and columns reveals the aggregate preference or predominance for certain interaction-algorithmic strategies among different age groups.
*Source:* Subfigure (a) adapted from Piatti et al. (2022).

## 4. Results

In this section, we present a detailed analysis of the outcomes derived from our evaluation of students' algorithmic skills using the Cross Array Task (CAT) in both unplugged and virtual formats, following the methodologies delineated in the previous section.

Hereafter, when discussing the interaction dimension (combination of artefact and autonomy), it is implied that we are referring to the lowest interaction dimension, which provides valuable insights into students' baseline competency levels.

### 4.1. Algorithmic thinking skills development

By examining students' interaction dynamics during the activity, focusing on the artefacts they used and their level of autonomy, we explore their impact on AT development, specifically on the algorithmic dimension achieved. This investigation shed light on effective strategies across both unplugged and digital domains. Additionally, we analyse how factors such as age, sex, and schema performance interact with these strategies and affect algorithmic complexity. We also examine how the interaction between age and sex influences algorithmic competence, emphasising the importance of understanding these differences in developing algorithmic skills.

#### 4.1.1. Analysis of interaction strategies

Fig. 2 illustrates students' interaction strategies and their impact on algorithmic complexity across different age groups, for both unplugged and virtual environments.

In the unplugged setting (Fig. 2(a)), younger pupils mainly used medium-complexity interactions (VS), while older ones leaned towards more complex methods (V). Interestingly, even very young students demonstrated an ability to conceive complex algorithms, although simpler 1D and 0D algorithms were more common. With age, there was a shift towards more complex 2D algorithms, although 1D algorithms remained prevalent.

In the virtual domain (Fig. 2(b)), younger primarily used the simpler interactions (GF), progressing to the G interaction (7–9 years) and balanced use of all modes (10–13 years). The oldest group (14–16 years) excelled in complex interactions (PF and P). Regardless of age, a common tendency was to create mainly 1D algorithms, with a developmental progression towards more complex 2D algorithms. Even the youngest age category demonstrated proficiency in complex 2D algorithms, surpassing simpler 0D ones. This highlights the early capability of young pupils to conceive intricate algorithms, especially in the virtual domain.

Overall, Fig. 2 reveals that as students mature, they progressively engage with more sophisticated interaction dimensions, shifting to voice-based in the unplugged environment and to programming in the virtual setting. The complexity of the algorithms created by students also escalates with age, with the virtual environment seemingly facilitating the use of more complex algorithms from an earlier age.

#### 4.1.2. Analysis of interaction strategies effect on algorithmic skills

To determine whether the interaction dimension is a predictor of the algorithm dimension and eventually determine if certain interaction strategies are more effective in producing complex algorithms, we performed a comprehensive statistical analysis for each domain.

ANOVA tests showed that artefact dimension is a significant predictor of algorithmic dimension in both virtual ($p < 1e-15^{****}$) and unplugged ($p < 1e-11^{****}$) domains.

T-tests reinforced these findings, revealing a considerable difference in algorithmic capability favouring the virtual setting ($t = -10.25365$, $p < 1e-23$).

Further analysis using Tukey's HSD test with BH adjustment and pairwise t-tests with Bonferroni adjustment highlighted a substantial impact of the PF artefact on algorithm dimension. Notably, the mean difference between PF and VSF was approximately 0.83 ($p < 1e{-}4^{****}$), while other comparisons also yield highly significant differences ($p < 0.001^{***}$) when PF is compared to the other artefacts.

The chi-squared test of proportions test added further weight to these findings, demonstrating highly significant variations in the prevalence of higher algorithm dimensions across different artefact categories ($\chi^2 = 140.38$, $p < 1e-15^{****}$). Notably, the virtual PF artefact exhibited the highest proportion of higher algorithm dimensions (46%), followed by P (39%), showcasing its remarkable efficacy in fostering more complex algorithmic constructs, particularly compared to the unplugged alternatives.

#### 4.1.3. Interaction strategies development in the virtual CAT

Focusing on the virtual CAT sample, central to our previous conclusion on the superiority of virtual artefacts, we analysed the interaction dimensions frequency, comparing the least complex ones with the most prevalent across age groups to discern variations in usage patterns. Fig. 3 illustrates a trend suggesting an interesting evolution in students' interaction preferences as pupils grow older.

In the youngest age category, where exposure to technology is limited, we restricted them to using programming interfaces, specifically allowing only gesture interactions. In this group, both the lowest and prevalent interaction dimensions align closely, with GF, the less complex option, being the predominant choice.

As students aged into the 7-9-year bracket, restrictions were lifted. While gesture interfaces remain the more popular assignment, some pupils explored visual programming interfaces, indicating a willingness to explore advanced methods.

Students in the third age group fairly used all four interaction dimensions, showcasing increased versatility and adaptability.

The oldest students leaned towards the most complex interaction dimensions (PF and P), preferring complexity and proficiency with more advanced methods.

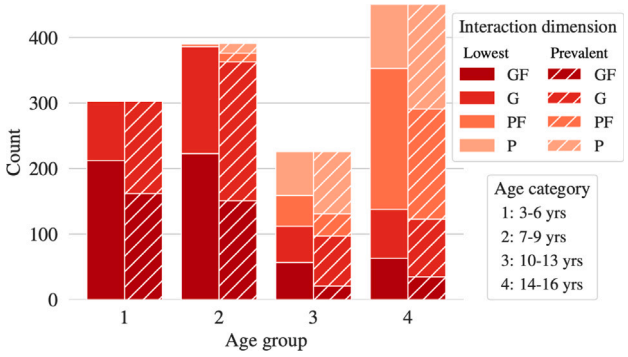



**Fig. 3. Distribution of interaction dimensions by age group in the virtual CAT.** The *y*-axis displays counts of the lowest and most used interaction dimensions, represented by solid and striped bars, across four age categories on the *x*-axis. The younger age group exclusively used the gesture interface (GF and G), as visual programming (PF and P) was not allowed.

#### 4.1.4. Analysis of age-related development of algorithmic skills

The ANOVA tests aimed at understanding the role of age in the development of AT, demonstrated a positive correlation between age and algorithmic complexity in the unplugged ($p < 1e{-}1^{****}$) and virtual ($p < 1e-6^{****}$) contexts, indicating that as the age increases, there is a corresponding increase in the complexity of the algorithm produced.

In the unplugged setting, older age groups (10–13 and 14–16) showed higher algorithmic dimensions and positive coefficients (0.23052, $p < 1e{-}06^{****}$; 0.36585, $p < 1e{-}14^{****}$). Similar patterns were observed in the virtual context for the older age category (0.23200, $1e{-}06^{***}$), reinforcing the link between older age categories and higher algorithm dimensions.

The chi-squared tests of proportions reveal significant age-related variations in higher algorithm dimensions in both domains ($p < 0.0001^{****}$), with proportions increasing with age, reaching 26% in the unplugged and 36% in the virtual environment for the 14 to 16 years old category.

#### 4.1.5. Analysis of the interplay of interaction strategies and age and their impact on algorithmic skills

To further detail the analysis, we examined the interaction between interaction strategies and age, demonstrating its significant effect on algorithmic skills. Interestingly, this interaction was found to be significant in both the unplugged ($p = 0.00839^{**}$) and virtual ($p = 0.000106^{***}$) environment, indicating that the effect of artefacts varies across age categories.

The linear regression model and the EMMs analysis yielded insightful distinctions in how different artefacts influence the algorithm dimension across various age groups. In the unplugged setting, artefact V was effective (0.56991, $p = 1e{-}06^{****}$), particularly for more mature learners (EMM = 1.060, SE = 0.0731), but its effectiveness diminished for younger age groups, where there is a marked preference for the simplest artefact, VSF (EMM = 1.500, SE = 0.4233). Conversely, in the virtual environment, the influence of artefact complexity on algorithm dimension appeared less directly correlated with age. Younger students benefited from GF (EMM = 1.090, SE = 0.0394), while older age groups engaged more with complex artefacts like PF (EMM = 1.465, SE = 0.0391) and P (EMM = 1.431, SE = 0.0409).

The chi-squared test of proportions across different age categories and artefacts yields intriguing insights into the distribution of algorithm dimensions. Significant differences were observed in each age group, with the most pronounced disparities in the oldest age category ($\chi^2 = 81.434$, $p < 1e-15^{****}$).

Dominant artefacts showed intriguing trends. In the unplugged setting, V is dominant for children aged 3 to 6 years, while VS takes over (15%) as they grow older, and V resurfaces as the dominant artefact (20%) for those aged 10 to 13 years and remains strong for 14 to 16-year-olds (26%). In the virtual domain, younger age groups prefer GF (24% and 38%), but a notable shift occurs at ages 10 to 13, where PF becomes dominant (45%), persisting for 14 to 16-year-olds (47%), closely followed by P (44%). These changes in dominant artefacts reveal the influential role of virtual domain artefacts, particularly PF, in shaping algorithm dimensions across diverse age groups.

#### 4.1.6. Analysis of the interplay of interaction strategies and sex and their impact on algorithmic skills

Following, we examined how interaction strategies and sex interact to influence algorithmic complexity. Our goal was to understand whether the interaction between these factors significantly affects the algorithmic dimension and whether these effects differ between the types of interaction performed.

The ANOVA reports contrasting results for the two environments. In particular, the interaction between artefact and sex is not significant in the unplugged domain ($p = 0.2655$), while in the virtual domain, this interplay is marginally significant ($p = 0.0521$).

The linear regression analysis confirms the finding from ANOVA. The significant interaction between the simplest virtual artefact G and sex (0.21394, $p = 0.00642^{**}$) suggests that the negative impact of artefact G on algorithmic complexity is less pronounced for males. This means that while artefact G generally reduces algorithmic complexity, this reduction is smaller for males than females.

**Table 4**
**Student participation and success rates across schemas.** The number and percentage of students who attempted and successfully completed each schema. Success rate is reported only for the virtual CAT, calculated from the number of students who attempted the schema. Rows shaded in grey indicate schemas with success rate exceeding 80%.

**(a) Unplugged CAT.**

| | | No. pupils total | No. pupils participating | **Participation** (%) |
|---|---|---|---|---|
| **Schema** | 1 | 109 | 109 | 100% |
| | 2 | 109 | 109 | 100% |
| | 3 | 109 | 109 | 100% |
| | 4 | 109 | 109 | 100% |
| | 5 | 109 | 109 | 100% |
| | 6 | 109 | 109 | 100% |
| | 7 | 109 | 109 | 100% |
| | 8 | 109 | 107 | 98% |
| | 9 | 109 | 105 | 96% |
| | 10 | 109 | 105 | 96% |
| | 11 | 109 | 105 | 96% |
| | 12 | 109 | 104 | 95% |

**(b) Virtual CAT**

| | | No. pupils total | No. pupils participating | **Participation** (%) | No. pupils succeeding | **Success** (%) |
|---|---|---|---|---|---|---|
| **Schema** | 1 | 129 | 126 | 98% | 119 | 94% |
| | 2 | 129 | 127 | 98% | 93 | 73% |
| | 3 | 129 | 127 | 98% | 100 | 79% |
| | 4 | 129 | 129 | 100% | 106 | 82% |
| | 5 | 129 | 128 | 99% | 109 | 85% |
| | 6 | 129 | 127 | 98% | 112 | 88% |
| | 7 | 129 | 125 | 97% | 98 | 78% |
| | 8 | 129 | 126 | 98% | 92 | 73% |
| | 9 | 129 | 121 | 94% | 98 | 81% |
| | 10 | 129 | 118 | 91% | 91 | 77% |
| | 11 | 129 | 110 | 85% | 82 | 75% |
| | 12 | 129 | 110 | 85% | 78 | 71% |

#### *4.1.7. Analysis of the interplay of interaction strategies and schemas and their impact on algorithmic skills*

By exploring how different interaction strategies and schemas interact to affect algorithmic skills, we aim to understand the joint influence of these factors on the algorithmic dimension and discuss possible reasons why the schema used might alter the effect of the artefact on algorithmic complexity.

ANOVA results reveal a significant interaction between these factors in the unplugged domain ($p = 1e-9^{****}$) and even stronger in the virtual one ($p = 1e-13^{****}$), suggesting a notable variation in how different artefacts affect algorithmic complexity depending on the schema performed.

In the unplugged environment, linear regression analysis identifies significant interactions between artefacts and specific schemas. For instance, artefact VS in combination with schemas 3, 4, 7, 8 and 12, as well as artefact PF with schemas 9, 10, 11, and 12, shows significant effects. The negative estimates for these interactions suggest that these strategies substantially reduce algorithmic complexity when applied to the given schemas. In the virtual environment, the analysis also highlights significant interactions. Artefact P, combined with schemas 2, 3, 8, 9, 10, 11, and 12, and artefact PF with schemas 9, 10, 11, and 12, shows significant effects. The positive estimates indicate that students applying these interaction strategies on specific schemas show higher algorithmic skills.

Finally, to understand whether and how the relationship between sex and algorithmic complexity varies by age, we examine how these factors interact and impact algorithmic skills. The significant interactions between age and sex on algorithmic skills in both environments suggest that the relationship between sex and algorithmic dimension is moderated by age.

ANOVA results show a significant interaction between age category and sex for both the unplugged environment ($p = 0.00364^{**}$) and the virtual one ($p = 0.000115^{***}$), indicating that the effect of sex on algorithmic skills differs by age.

The linear regression model for the unplugged environment reveals that male participants aged 10 to 13 years show a higher algorithmic dimension compared to their female counterparts (0.22245, $p = 0.011^{*}$), while the interaction for older age groups is not significant. This indicates that, for this age range, sex differences in algorithmic skills are more pronounced. In the virtual environment, the situation is reversed. Male participants aged 10 to 13 years exhibit lower algorithmic dimensions compared to females, reflecting a potential sex disparity that reverses from the unplugged environment (−0.29045, $p = 0.00641$).

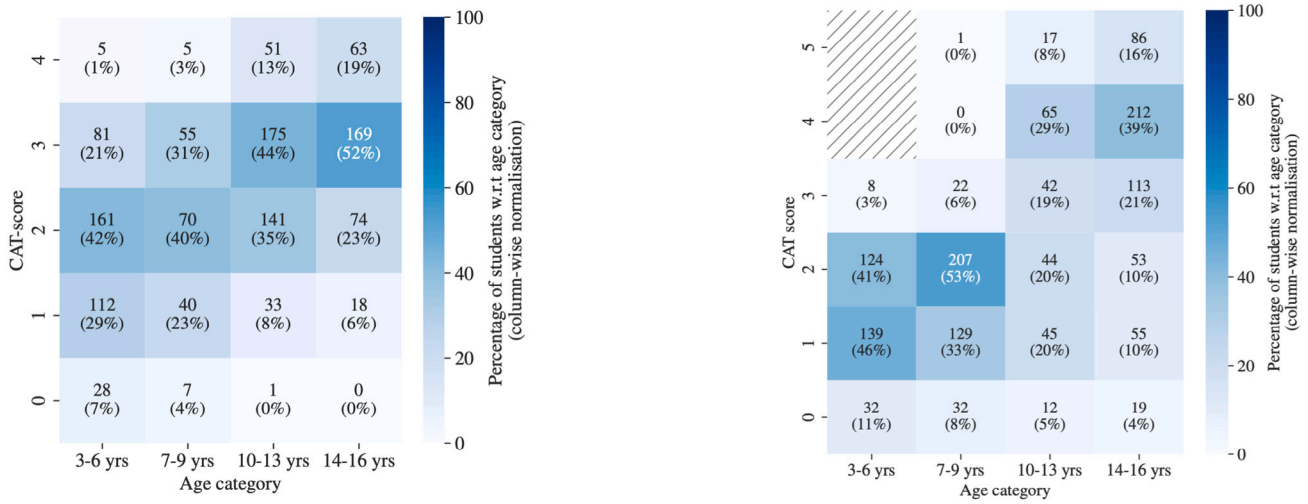
### *4.2. Student participation and success*

#### *4.2.1. Analysis of the interplay of age and sex and their impact on algorithmic skills*

Table 4 provides a detailed overview of student participation and success on individual schemas. On the one hand, for the unplugged CAT, pupils successfully tackled tasks up to schema 7. Beyond that point, some experiments were heated due to time constraints, though the participation rate remained high, with a minimum of 95%. On the other hand, for the virtual CAT, the participation rate, while still notably high, experienced a slight decrease, reaching a minimum of 85%. This decrease could be attributed to students having the autonomy to interrupt their participation voluntarily, something they could not do in the unplugged CAT. Therefore, high participation in the virtual CAT indicates strong determination and intrinsic interest. Regarding success rates, in the unplugged CAT, all students successfully completed each schema they attempted, as they were guided to correct any errors they made. However, for the virtual CAT, success rates varied. Schema 1 had the highest success rate at 94%, fitting its role as an easy introductory task. Conversely, schemas 11 and 12 show success rates of 75% and 71%, indicating increased complexity. The drop in success rates for schemas 2 and 8 to 73%, despite similar attempts to schema 1, may indicate heightened task complexity or a mismatch between students' skills and schema demands. The non-linear decline in success

**Table 5**
**Student participation and success rates across age categories in the virtual CAT.** The number and percentage of students who attempted and successfully completed all 12 schemas, grouped by age category, along with the median and range of schemas attempted and succeeded, with the interquartile range (IQR) indicated by the median (Q2) and spanning from the 25th percentile (Q1) to the 75th percentile (Q3). Success rate is calculated from the number of students participating.

| | | No. pupils total | No. pupils participating in all 12 schemas | **Participation** (%) | Median schemas participated (Q1-Q3) | No. pupils succeeding in all 12 schemas | **Success** (%) | Median schemas succeeded (Q1-Q3) |
|---|---|---|---|---|---|---|---|---|
| **Age category** | 3-6 yrs | 27 | 18 | 67% | 12 (11-12) | 6 | 33% | 10 (8-11) |
| | 7-9 yrs | 33 | 30 | 91% | 12 (12-12) | 4 | 13% | 10 (9-11) |
| | 10-13 yrs | 20 | 13 | 65% | 12 (11-12) | 3 | 23% | 10 (7-11) |
| | 14-16 yrs | 49 | 38 | 78% | 12 (12-12) | 11 | 29% | 10 (8-11) |




(a) Unplugged CAT. (b) Virtual CAT.

**Fig. 4. Age-wise distribution of CAT score levels.** The *y*-axis represents CAT score levels, the *x*-axis depicts age categories, and each cell displays the number and percentage of students falling into specific score level categories.

rates as the schema numbers increase suggests that students perceive varying difficulty levels, which may not always align with educators' intended challenges. This is confirmed by the REML model analysis in Section 4.4.5, highlighting variations in students' experiences of difficulty within a set of tasks that, contrary to initial assumptions, do not follow a gradual order of difficulty, in line with the unplugged CAT results (Piatti et al., 2022).

Expanding the analysis to the virtual domain, Table 5 illustrates trends in student participation and success across age categories. Participation rates exceeded 65% for all age groups, with 7–9 year-olds showing a particularly high attempt rate of 91%. In contrast, success rates revealed that the youngest and oldest students achieved the highest success rates. Students clustered within a consistent range of solved schemas in each age category, a part of those in the 10–13 category had a slightly wider interquartile range (IQR), suggesting that they might have attempted fewer schemas than other age groups.

Examining student performance across age groups in Fig. 4 for both the unplugged and virtual CAT, it is evident that performance increases with age. These differences are statistically significant for both the unplugged ($\chi^2 = 276.21$, $p < 1e-15^{****}$) and virtual ($\chi^2 = 735.73$, $p < 1e-15^{****}$) domains. The result from the Tukey HSD for pairwise comparisons in Table 6 shows that the significance holds for all age categories except the two youngest ones, where the unplugged domain shows a decrease in significance, and the virtual domain lacks significance.

Returning to the virtual domain, the tendencies across the interaction dimensions, illustrated in Table 7, indicate participation levels ranging from 39% to 67%, with G and GF interactions leading at 67% and 66%, respectively. GF stands out as the interaction dimension with the highest median and a more consistent pattern of attempting a greater number of schemas. Success rates consistently range from 90% to a perfect 100%, with PF showing a perfect score. While high in participation, G exhibits a lower success rate than programming interfaces (PF and P), suggesting limitations in handling complex tasks. Notably, artefacts with feedback (GF and PF) exhibit higher success rates, with GF having the highest median success rate and a moderate spread, and PF, though not reaching the same median success rate, demonstrates effectiveness with a moderate spread in individual success rates.

**Table 6**
**Pairwise comparison between age groups.** The result from Tukey's HSD test for pairwise comparisons with Benjamini–Hochberg *p* value correction for false rate detection rates.

**(a) Unplugged CAT**

| | 3-6 yrs | 7-9 yrs | 10-13 yrs |
|---|---|---|---|
| 7–9 yrs | $p < 0.01^{**}$ | | |
| 10–13 yrs | $p < 0.0001^{****}$ | $p < 0.0001^{****}$ | |
| 14–16 yrs | $p < 0.0001^{****}$ | $p < 0.0001^{****}$ | $p < 0.001^{***}$ |

**(b) Virtual CAT**

| | 3-6 yrs | 7-9 yrs | 10-13 yrs |
|---|---|---|---|
| 7–9 yrs | $p = 0.0542$ | | |
| 10–13 yrs | $p < 0.0001^{****}$ | $p < 0.0001^{****}$ | |
| 14–16 yrs | $p < 0.0001^{****}$ | $p < 0.0001^{****}$ | $p < 0.0001^{****}$ |

$^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$, $^{****}\ p < 0.0001$

**Table 7**
**Student participation and success rates across interaction dimensions in the virtual CAT.** The number and percentage of students who attempted and successfully completed any schema, grouped by interaction dimension, along with the median and range of schemas attempted and succeeded, with the interquartile range (IQR) indicated by the median (Q2) and spanning from the 25th percentile (Q1) to the 75th percentile (Q3). Success rate is calculated from the number of students participating.

| | | No. pupils total | No. pupils participating in any schema | **Participation** (%) | Median schemas participated (Q1-Q3) | No. pupils succeeding in any schemas | **Success** (%) | Median schemas succeeded (Q1-Q3) |
|---|---|---|---|---|---|---|---|---|
| **Interaction dimension** | GF | 129 | 85 | 66% | 7 (4-9) | 78 | (92%) | 6 (3-8) |
| | G | 129 | 87 | 67% | 4 (1-7) | 79 | (91%) | 2 (1-4) |
| | PF | 129 | 50 | 39% | 5 (2-8) | 50 | (100%) | 4 (2-7) |
| | P | 129 | 58 | 45% | 4 (1-7) | 52 | (90%) | 3 (1-6) |

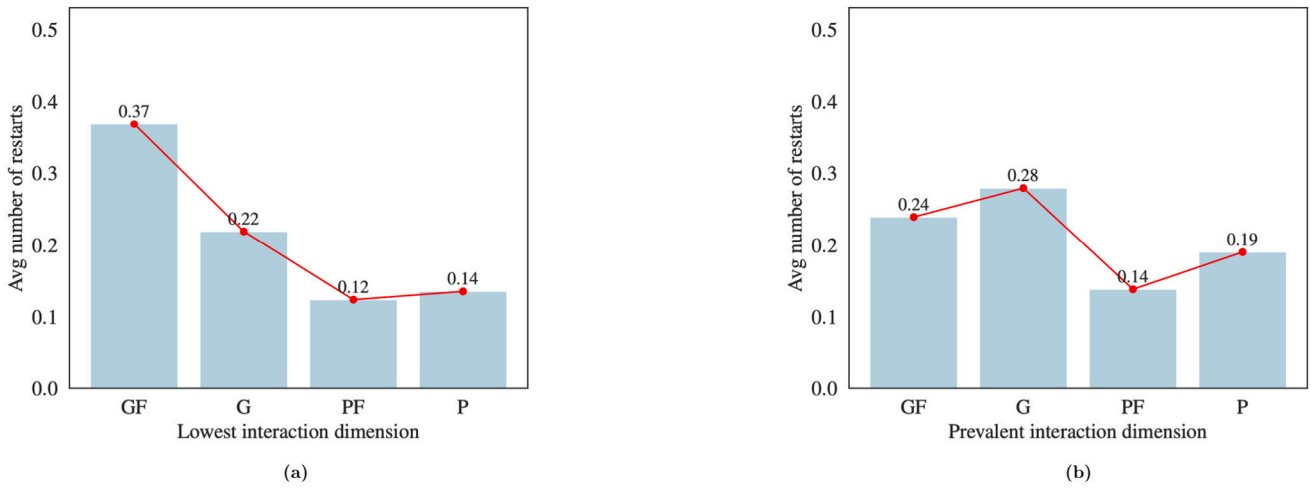



**Fig. 5. Restarts distribution per interaction dimension.** For each interaction dimension on the $x$-axis, specifically in (a) representing the lower-level interaction dimension used, and in (b) representing the most frequently used interaction dimension, the $y$-axis displays the average number of restarts.

### 4.3. Trial and error strategies

The analysis of task restart frequency, as an indicator of trial and error (T&E) strategies, provides valuable insights into the factors influencing this behaviour and its impact on performance outcomes.

#### 4.3.1. Factors influencing restarts

Our findings indicate that neither the characteristics and complexities of schemas (−0.0098, $p = 0.161$) nor sex (−0.0059, $p = 0.902$), significantly influence restart behaviour.

Exploring the interaction dimension's influence, a negative correlation between complexity and restart frequency is observed, indicating that as students work with increasingly complex artefacts, they tend to restart their tasks less frequently, as shown by the coefficients for both the lowest interaction dimension (−0.0283, $p = 0.180$) and the prevalent interaction dimension (−0.0168, $p = 0.427$). Fig. 5(a) supports this, showing that, for the lowest artefact used, the average number of restarts decreases when dealing with more complex artefacts. Fig. 5(b) further shows a non-linear relationship between prevalent artefact complexity and restarts. The slight increase in restarts from non-autonomous (GF and PF) to autonomous use (G and P) suggests that visual feedback within artefacts is crucial in supporting students' task participation and reducing restarts, possibly indicating uncertainty when such support is absent.

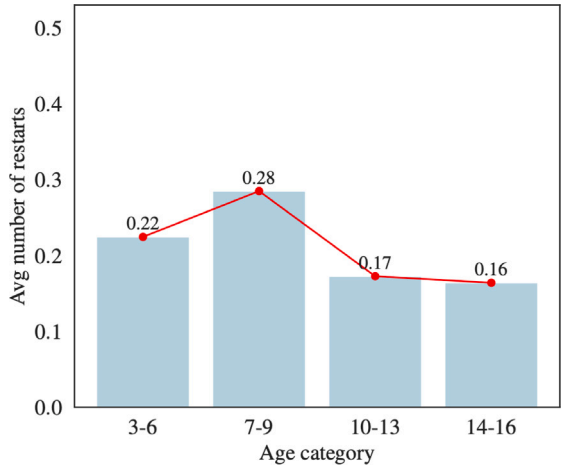



**Fig. 6. Restarts distribution per age.** For each age category on the $x$-axis, the average number of restarts is reported on the $y$-axis.

Age plays a significant role in restart behaviour, as reflected in a statistically significant inverse relationship (−0.0167, $p = 0.012^{*}$), indicating that older students are less inclined to restart tasks. Fig. 6 illustrates this decrease in the average number of restarts with increasing age categories. A peak in restarts among the 7–9 age group suggests increased exploration or developing problem-solving efficiency, while the decline in restarts among older groups (10–13 and 14–16 years) may signal improved problem-solving skills and a greater ability to integrate past experiences, indicating a shift towards more advanced problem-solving approaches as students mature.

#### 4.3.2. Restarts influence on performance

Our regression analysis further probed the relationship between restart behaviour and performance outcomes, including algorithm complexity and CAT scores, yielding contrasting results. For the algorithm dimension, there is no significant relationship with restarts (−0.0221, $p = 0.232$), indicating that restart frequency may not reliably predict algorithmic performance. Similarly, the link between restarts and CAT scores was not statistically significant (−0.0677, $p = 0.110$).

Our exploration of potential non-linear relationships, using polynomial terms in regression models, did not reveal significant patterns for algorithmic performance ($p = 0.200, 0.407$, and 0.497).

In contrast, the analysis of CAT scores with polynomial terms uncovered a nuanced relationship. Initially, increased restarts were associated with decreased scores (−0.4509, $p = 0.001^{**}$). However, introducing a quadratic term revealed a positive effect (0.0895, $p = 0.031^{*}$), suggesting a non-linear connection, indicating that beyond a certain threshold, an increase in restarts leads to improvements. Despite this, the cubic term did not achieve statistical significance (−0.0037, $p = 0.151$), indicating that additional complexity did not enhance our understanding of this relationship.

### 4.4. Linear mixed model assessment of student performance

#### 4.4.1. Model selection and refinement

Analysing the baseline model (M0), defined in Section 3.4.4, from the model summary in Table 8 indicates slight non-significant variations in CAT scores between male and female students, with a negative coefficient suggesting slightly lower performance in males ($p = 0.497$). Similarly, regarding canton, students from Ticino do not show statistically significant differences in CAT scores compared to students from Solothurn ($p = 0.719$). Variance estimates reveal variation in student performance across different grouping levels, particularly at the school and student levels.

**Table 8**
**Baseline model summary.** The REML criterion at convergence is 4044.3 for the baseline linear mixed-effects model.

**(a) Scaled residuals.**

| Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|
| −3.785 | −0.582 | 0.050 | 0.611 | 3.569 |

**(b) Random effects.** Number of observations: 1457.
Groups: Student, 129; Schema, 12; Session-Grade, 9; School, 5.

| Groups | Name | Variance | *SD*[a] |
|---|---|---|---|
| Student | (Intercept) | 0.490 | 0.700 |
| Schema | (Intercept) | 0.057 | 0.238 |
| Session-Grade | (Intercept) | 0.054 | 0.232 |
| School | (Intercept) | 0.931 | 0.965 |
| Residual | | 0.756 | 0.870 |

[a] Standard deviation

**(c) Fixed effects.**

| | Estimate | *SE*[b] | df | *t*[c] | *p*[d] |
|---|---|---|---|---|---|
| (Intercept) | 2.169 | 0.986 | 2.826 | 2.200 | 0.121 |
| Sex[e] | −0.090 | 0.132 | 120.082 | −0.681 | 0.497 |
| Canton[f] | 0.439 | 1.101 | 2.815 | 0.398 | 0.719 |

[b] Standard error
[c] $t$ value
[d] $p$ value $= Pr(>|t|)$
[e] Sex: Male
[f] Canton: Ticino

**Table 9**
**LRT to evaluate the inclusion of canton as a predictor.** Comparison between the reduced model (M1) without the canton predictor and the baseline model (M0) including it.

| Model | AIC[a] | $\chi^2$ | *p*[b] |
|---|---|---|---|
| M1 | 4058.1 | | |
| M0 | 4059.8 | 0.275 | 0.5998 |

[a] Akaike Information Criterion for the model evaluated as $-2 \cdot (\text{logLik} - \text{npar})$. Smaller is better.
[b] $p$ value $= Pr(> \chi^2)$.

We proceed by iteratively comparing and refining our model, starting with the baseline (M0), better to understand the predictors' contributions to student performance. The LRT results in Table 9 indicate that including the canton predictor does not significantly improve the model's fit ($p = 0.5998$), likely due to the limited representation of Swiss cantons in the data (only 2 out of 26 sampled). Therefore, we opt for the simpler reduced model (M1). To assess the impact of sex, whose effect was not statistically significant, we compared three models by considering different combinations of predictors. The LRT results in Table 10 show that including sex as a fixed effect (M2) did not significantly enhance model fit ($p = 0.4881$), suggesting it may not be a substantial predictor of CAT scores alone. However, the improved model (M3) without the predictor shows a significant improvement in fit compared to the model without sex ($p < 1e-3^{***}$), indicating that sex-related differences may vary across schools.

$$\begin{aligned} \texttt{CAT\_SCORE} = &\beta_0 + (\beta_1 + \beta_2 \cdot \texttt{SEX} + \epsilon_{\texttt{SEX}} | \texttt{SCHOOL}) + \\ &+ u_{\texttt{STUDENT}} + u_{\texttt{SESSION\_GRADE}} + u_{\texttt{SCHEMA}} + \epsilon. \end{aligned} \tag{3}$$

Our chosen model (M3), defined in (3), comprises $\beta_0$, which represents the intercept coefficient, while $\beta_1$ and $\beta_2$ are the coefficients for the effect of sex on school. As random effect we included `STUDENT`, `SESSION_GRADE`, and `SCHEMA`. Finally, $\epsilon_{\text{SEX}}$ is the error term for the interaction between sex and school, while $\epsilon$ is the unexplained variability in the model.

The summary of the models is provided in Table 11. The intercept in the fixed effects fis significant for both the unplugged ($p = 0.014^{*}$) and virtual domains ($p = 0.00286^{**}$). This suggests that the average CAT score significantly differs from zero when accounting for random effects, indicating a baseline proficiency level within the student population. Exploring the random effects, in both domains the largest source of variance is attributed to school-level differences (U = 0.47939, V = 0.87332), suggesting that factors unique to each school significantly influence CAT scores. The variance related to sex is similar for both domains, with a lower positive correlation in the unplugged domain (0.14) compared to the negative in the virtual domain (−0.49). This indicates that sex dynamics and school-specific factors impact male and female students' performance differently, with a trend in the virtual CAT where schools with higher overall performance may have lower scores for male students and vice versa. The variance associated with individual students is higher for the virtual domain (U = 0.23476, V = 0.40687). This difference possibly reflects the diversity in students' abilities and the influence of other unmeasured factors, with a more homogeneous response observed with unplugged artefacts.

**Table 10**
**LRT to evaluate the inclusion of sex as a predictor.** Comparison between the reduced model (M1) without the canton predictor, another reduced model (M2) without canton and sex, and an improved model (M3) without canton, but that considers sex as a random slope within schools. Rows shaded in grey indicate statistically significant models.

| Model | AIC[a] | $\chi^2$ | *p*[b] |
|---|---|---|---|
| M2 | 4056.6 | | |
| M1 | 4058.1 | 0.481 | 0.4881 |
| M3 | 4048.3 | 11.785 | 0.0006*** |

$^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$, $^{****}$ $p < 0.0001$
[a] Akaike Information Criterion for the model evaluated as $-2 \cdot (\text{logLik} - \text{npar})$. Smaller is better.
[b] $p$ value $= Pr(> \chi^2)$

**Table 11**
**Model summary.** The REML criterion at convergence is 3377.3 for the unplugged domain (U) and 4032.5 for the virtual domain (V) in the linear mixed-effects models.

**(a) Scaled residuals.**

| | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|
| U | -3.574 | -0.656 | -0.048 | 0.563 | 2.855 |
| V | -3.797 | -0.578 | 0.048 | 0.616 | 3.537 |

**(b) Random effects.** The unplugged domain (U) has 1280 observations with groups: Student (109), Schema (12), Session-Grade (8), School (3). The virtual domain (V) has 1457 observations with groups: Student (129), Schema (12), Session-Grade (9), School (5).

| | Groups | Name | Variance | *SD*[a] | Corr |
|---|---|---|---|---|---|
| U | Student | (Intercept) | 0.235 | 0.485 | |
| | Schema | (Intercept) | 0.197 | 0.443 | |
| | Session-Grade | (Intercept) | 0.024 | 0.154 | |
| | School | (Intercept) | 0.479 | 0.692 | |
| | | Sex[b] | 0.027 | 0.164 | 0.14 |
| | Residual | | 0.671 | 0.819 | |
| V | Student | (Intercept) | 0.407 | 0.638 | |
| | Schema | (Intercept) | 0.057 | 0.237 | |
| | Session-Grade | (Intercept) | 0.059 | 0.243 | |
| | School | (Intercept) | 0.873 | 0.935 | |
| | | Sex[b] | 0.253 | 0.503 | -0.49 |
| | Residual | | 0.756 | 0.870 | |

[a] Standard deviation
[b] Sex: Male

**(c) Fixed effects.** Rows shaded in grey indicate statistically significant fixed effects.

| | | Estimate | *SE*[c] | df | *t*[d] | *p*[e] |
|---|---|---|---|---|---|---|
| U | (Intercept) | 2.827 | 0.432 | 2.397 | 6.54 | 0.014* |
| V | (Intercept) | 2.429 | 0.389 | 4.194 | 6.244 | 0.003** |

$^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$, $^{****}$ $p < 0.0001$
[c] Standard error
[d] $t$ value
[e] $p$ value $= Pr(>|t|)$

For session-grade, the unplugged dataset shows a smaller variance (U = 0.02378, V = 0.05883), suggesting more consistent performance across different sessions than the virtual domain. This may indicate a more stable impact of session-grade factors like age progression, curriculum complexity, teaching methods, or cohort effects on student performance. Finally, the significant difference in variability attributed to schemas between the unplugged and virtual domains (U = 0.19660, V = 0.05605) suggests that differences in activity nature, information presentation, or features like the ability to skip or solve schemas in preferred order may contribute to this variation.

**Table 12**
**Type III analysis of variance (ANOVA) table with Satterthwaite's method.** Rows shaded in grey indicate statistically significant variables.

| | | AIC[a] | LRT[b] | $p$[c] |
|---|---|---|---|---|
| | | 3393.3 | | |
| | Schema | 3653.2 | 261.903 | <1e-15**** |
| U | Sex[d] | 3389.7 | 0.438 | 0.803 |
| | Session-Grade | 3393.1 | 1.780 | 0.182 |
| | Student | 3584.7 | 193.418 | <1e-15**** |
| | | 4048.5 | | |
| | Schema | 4112.5 | 65.75 | <1e-15**** |
| V | Sex[d] | 4056.7 | 12.20 | 0.002** |
| | Session-Grade | 4049.6 | 3.11 | 0.078 |
| | Student | 4397.0 | 350.52 | <1e-15**** |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$
[a] Akaike Information Criterion for the model evaluated as $-2 \cdot (\text{logLik} - \text{npar})$. Smaller is better.
[b] LRT statistic; twice the difference in log-likelihood, which is asymptotically chi-square distributed.
[c] $p$ value = $Pr(> \chi^2)$
[d] Sex in (1 + Sex | School).

The ANOVA results in Table 12 further confirm the significance of the random effects, particularly for Schema and student ($p < 1e-15$****), highlighting their crucial role in the model's explanatory power and indicating that both the type of task and individual student differences have a substantial impact on CAT scores. Interestingly, the impact of sex on CAT scores varies significantly across schools in the virtual dataset ($p = 0.002$**), while in the unplugged dataset, this variation is not statistically significant ($p = 0.803$), indicating that the role of sex-related factors is more pronounced in a virtual environment, contributing to varying outcomes in student performance across different schools.

#### 4.4.2. Sex influence within schools on performance

Comparing sex-related school performance in the two domains, from Fig. 7, we see no significant sex effect across schools for the unplugged domain. In contrast, the virtual domain exhibits variability in CAT scores between male and female students across schools. Certain schools (e.g., D) show higher CAT scores for male students than others (e.g., F), challenging the idea of a uniform sex effect.

Focusing on average performances across schools, we discern differences in the baseline performance for both domains. In the unplugged case, A performs below average, B slightly below, and C notably above. In the virtual case, schools exhibit varied impacts on female students, with some (A, D, and E) showing decreased CAT scores and others (F and G) demonstrating increased scores.

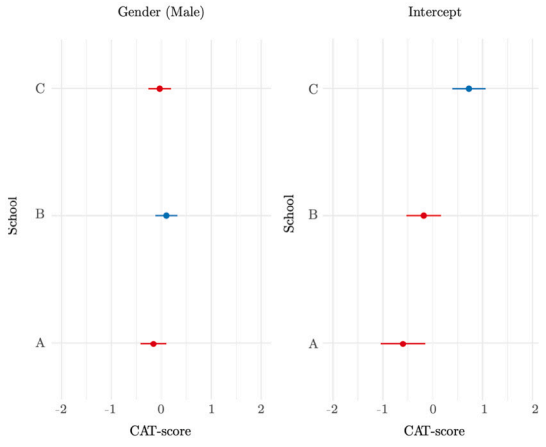
#### 4.4.3. Individual student variability in performance

The analysis of student performance in Fig. 8 showed significant individual variability in both domains. The presence of high achievers (blue dots to the right) and those facing challenges (red dots to the left) is consistent across both datasets, highlighting a diverse range of performances. This indicates a substantial amount of unexplained variability not accounted for by other factors considered in the study.

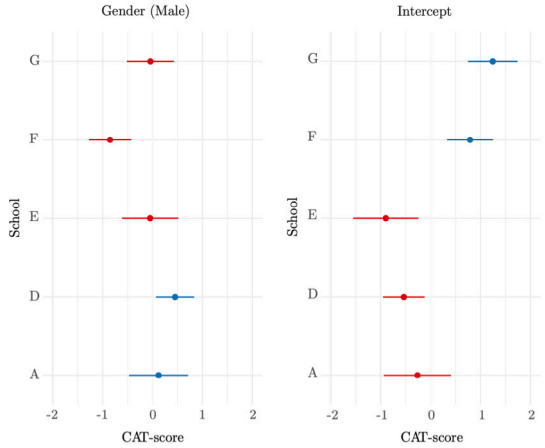
Despite accounting for school-level differences in the model, unexplained variability in CAT scores persists among students, indicating significant differences in performance residuals across various schools. This suggests that factors associated with the distinct educational environments of each school might contribute to the observed variance. Statistical analysis, specifically Levene's test, supports this observation ($p < 1e-15$****).

#### 4.4.4. Session grade impact on performance

Examining the impact of session and grade on scores in Fig. 9, no statistical differences in performance across sessions and grades are observed in both the unplugged and virtual domains. The pattern of fluctuations implies a complex relationship between sessions, grades, and CAT scores. Notably, lower performance is observed from the initial to the middle sessions, coinciding with lower HarmoS grades (HGs). Positive deviations in higher sessions suggest that older students generally perform better. This consistency implies that advanced cognitive skills and better adaptation to educational demands may contribute to improved performance among older students.

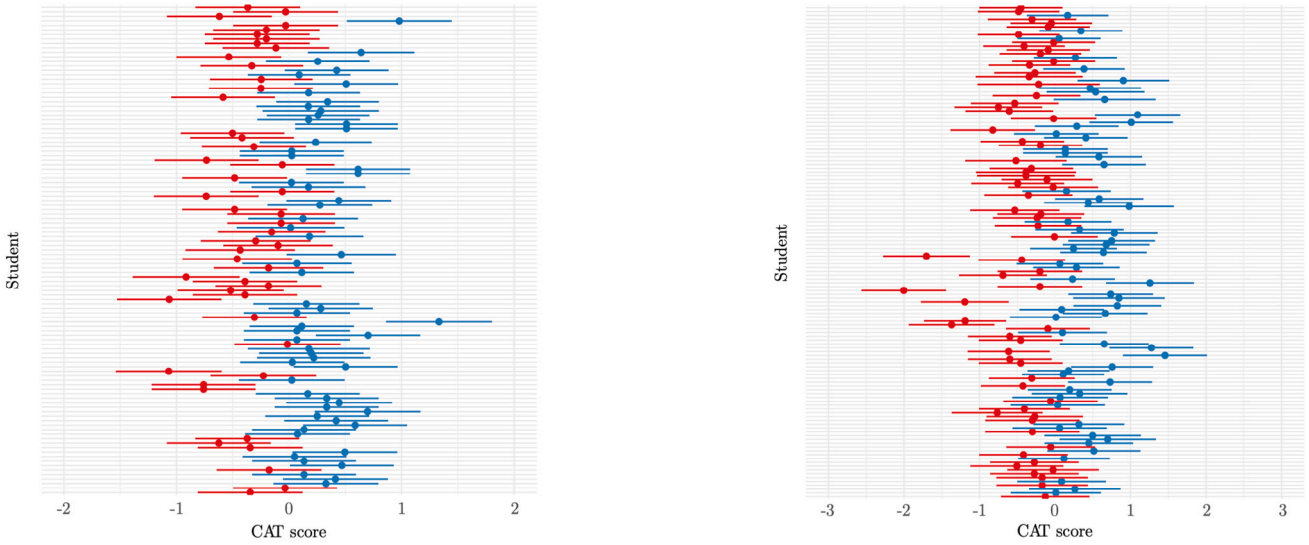



(a) Unplugged CAT.



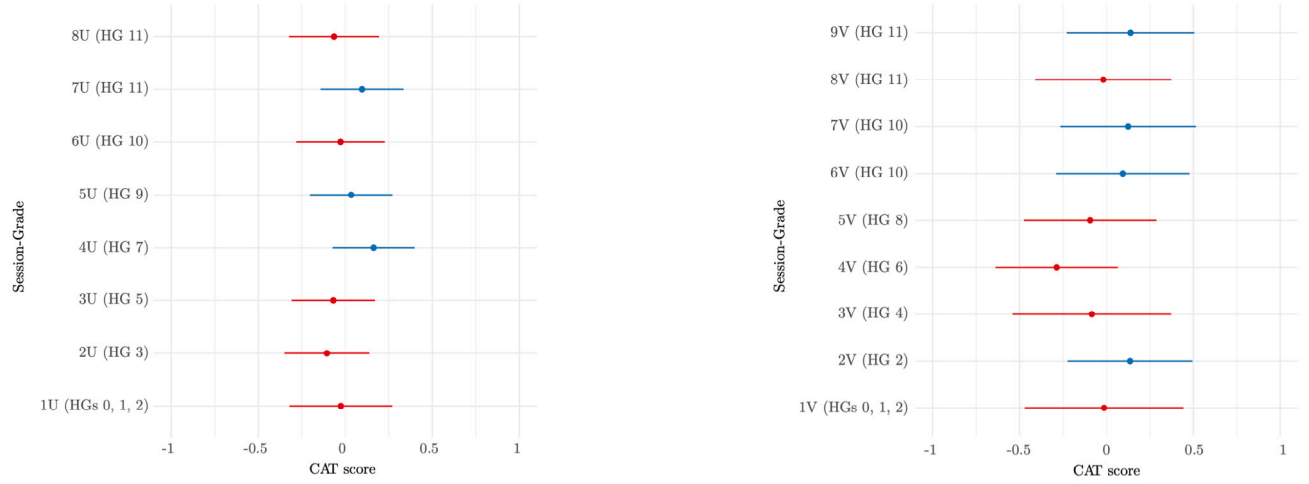

(b) Virtual CAT.

**Fig. 7. Sex-related school performance variations.** The plot on the left captures the variability in CAT scores between schools for male students compared to female students, illustrating how scores differ across schools based on sex. On the right, the plot shows the intercept, representing the average variability between schools, exclusively focusing on female students. Blue points represent scores above average, while reds those below. Horizontal lines represent the estimates' confidence intervals.




(a) Unplugged CAT. (b) Virtual CAT.

**Fig. 8. Individual student performance variations.** Each point represents the student deviation from the average CAT score, with blue indicating scores above average and red below. Horizontal lines represent the estimates' confidence intervals.




(a) Unplugged CAT. (b) Virtual CAT.

**Fig. 9. Session-Grade performance variations.** Each point represents the session deviation from the average CAT score, with blue indicating scores above average and red below. Horizontal lines represent the estimates' confidence intervals.

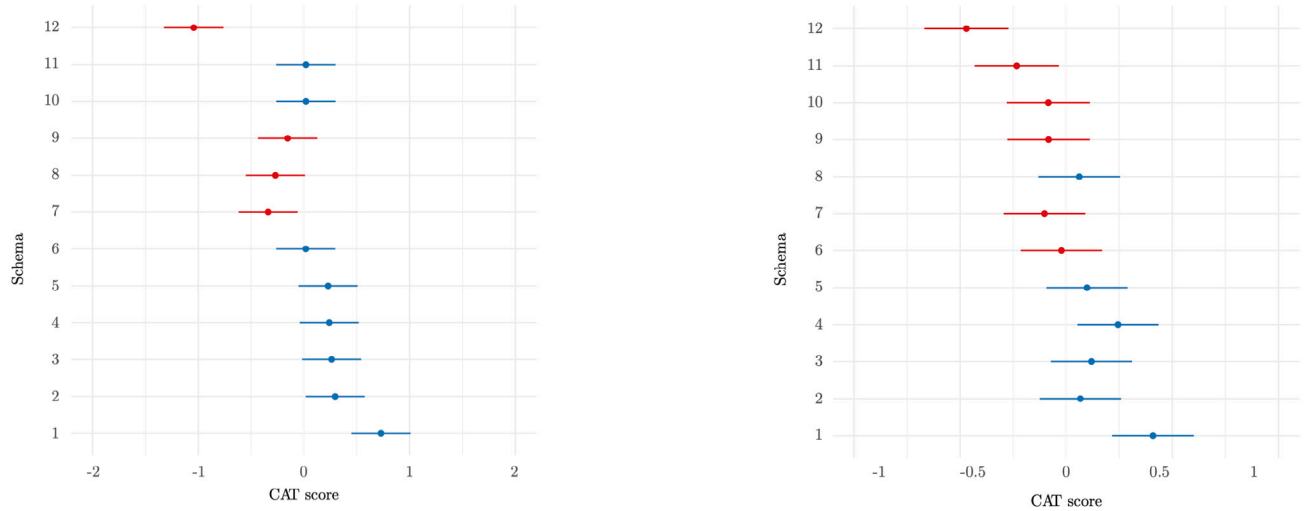
#### *4.4.5. Schema-based differences in performance*

Task performance varies across different schemas in both the unplugged and virtual domains, as depicted in Fig. 10. For the unplugged dataset, initial schemas (1 to 6) generally yield good performance, although there is a decline as the schema number increases, hinting at rising task difficulty. Schemas 7 to 9 show below-benchmark scores but with improving trends, suggesting student adaptation or better task alignment. Scores rise in schemas 10 and 11 but drop significantly in schema 12, possibly due to task difficulty or misalignment with student abilities. A consistent decreasing trend is observed in the virtual dataset, although with some irregularities. Performance is above the benchmark for less difficult tasks (1 to 5) and declines below the benchmark (6 to 12) with increasing task difficulty. Notably, the mean CAT score for schema 8 is above the benchmark, suggesting better-than-expected performance on average.

To explore performance trends and irregularities across different tasks, we specifically looked at the algorithm dimension instead of overall performance. This examination pertains specifically to the virtual CAT, where we have precise and comprehensive information on all the commands students use in crafting their algorithms. Fig. 11(a) indicates that the algorithm dimension varies across tasks, suggesting that students adapt their problem-solving strategies to each task rather than following a linear regression of algorithm complexity. Notably, for schemas 1, 2, 5, 6, and 12, students often use 1D dimensional algorithms driven by practical considerations. Sometimes, a simpler, less complex algorithm with fewer moves is more effective than a more intricate one. This preference for efficiency does not imply lower performance but reflects a pragmatic approach to problem-solving, as argued in Section 2.4.1. To assess student performance, we introduced an alternative method considering both algorithm complexity and efficiency. The adapted algorithm dimension metric, presented in Fig. 11(b), demonstrates a more linear decrease in average algorithm dimensions. Fig. 12(b) shows the original and updated distribution of performance across schemas using the new metric.

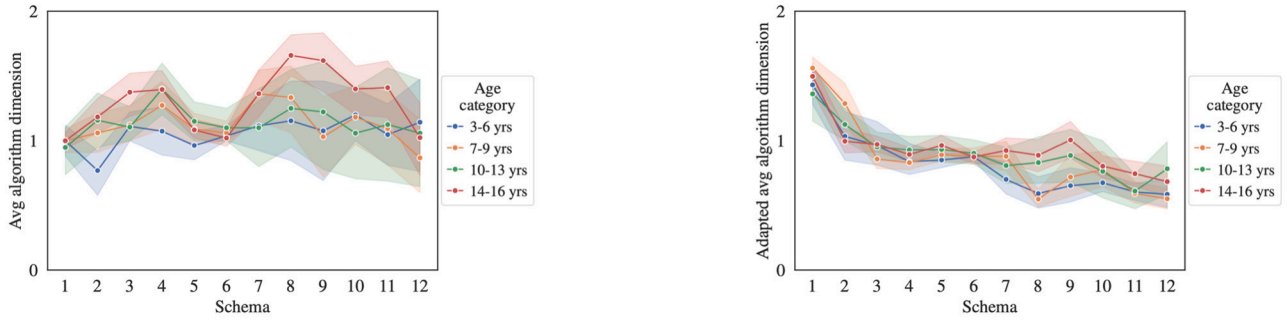
#### *4.4.6. Effect of task completion time on performance*

In the concluding phase of our analysis, we incorporated task completion time as a random effect into our existing model. This specific



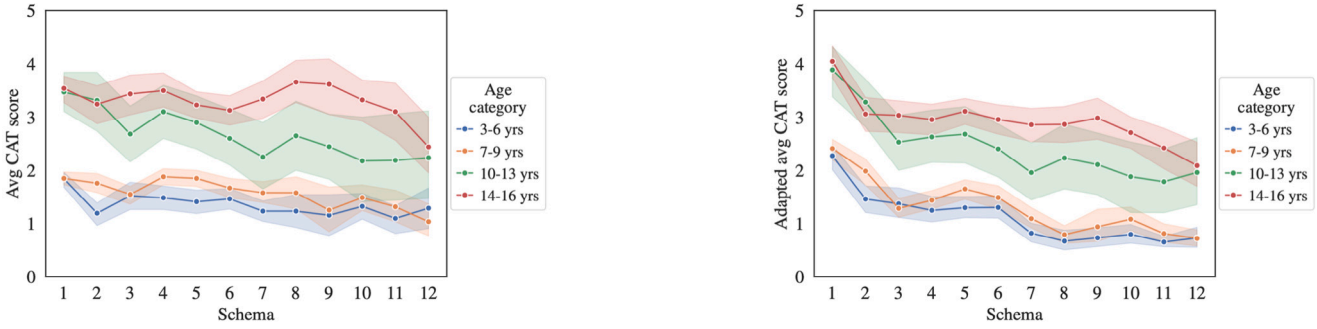

(a) Unplugged CAT. (b) Virtual CAT.

**Fig. 10. Schema-based performance variations.** Each point represents the schema deviation from the average CAT score, with blue indicating scores above average and red below. Horizontal lines represent the estimates' confidence intervals.




(a) Standard algorithm dimension metric. (b) Adapted algorithm dimension metric.

**Fig. 11. Algorithm dimension variations across age categories at schema level.** The *y*-axis represents the average variations in algorithm dimension for each age category, plotted against different schemas on the *x*-axis.




(a) Standard CAT score metric. (b) Adapted CAT score metric.

**Fig. 12. Performance variations across age categories at schema level.** The *y*-axis represents the average variations in CAT score for each age category, plotted against different schemas on the *x*-axis.

examination was exclusive to the virtual CAT, benefiting from detailed records of task completion times. We aimed to uncover the correlation between the time students spent on tasks and their ensuing performance levels.

In Fig. 13, a non-linear relationship between task completion time and performance is evident. Both extremely brief and significantly extended durations appear beneficial, resulting in higher CAT scores. This observation suggests that rapid responses may be driven by strong intuition or familiarity, while longer times may reflect a more analytical approach, likely enhancing performance. On the other hand, intermediate completion times do not seem to capitalise on the strengths of either approach, potentially explaining the observed dip in scores and the negative impact on performance associated with moderate haste.

**Table 13**
**Activity completion time across age categories.** The time spent by students to complete all 12 schemas, including mean, minimum and maximum values, median, and interquartile ranges (Q1-Q3), grouped by age category. The overall summary statistics at the bottom provide an overview of completion times for the entire dataset.

| | | Mean | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|---|---|
| **Age category** | 3–6 yrs | 20 m 53 s | 03 m 41 s | 15 m 48 s | 21 m 50 s | 27 m 14 s | 35 m 51 s |
| | 7–9 yrs | 13 m 13 s | 05 m 27 s | 09 m 52 s | 12 m 24 s | 13 m 54 s | 29 m 39 s |
| | 10–13 yrs | 26 m 41 s | 05 m 19 s | 17 m 04 s | 25 m 04 s | 41 m 51 s | 52 m 26 s |
| | 14–16 yrs | 29 m 34 s | 02 m 51 s | 18 m 58 s | 28 m 08 s | 40 m 29 s | 79 m 36 s |
| | | 23 m 07 s | 02 m 51 s | 12 m 14 s | 20 m 51 s | 30 m 03 s | 79 m 36 s |

**Table 14**
**Time spent using each interaction dimension.** The time students spent using a certain interaction dimension, including mean, minimum and maximum values, median, and interquartile ranges (Q1-Q3).

| | | Mean | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|---|---|
| **Interaction dimension** | GF | 11 m 38 s | 00 m 29 s | 05 m 21 s | 10 m 00 s | 16 m 59 s | 35 m 36 s |
| | G | 02 m 49 s | 00 m 04 s | 00 m 45 s | 01 m 49 s | 04 m 30 s | 11 m 52 s |
| | PF | 21 m 13 s | 02 m 31 s | 08 m 32 s | 18 m 20 s | 30 m 05 s | 79 m 36 s |
| | P | 12 m 15 s | 00 m 01 s | 00 m 29 s | 07 m 18 s | 19 m 33 s | 45 m 20 s |

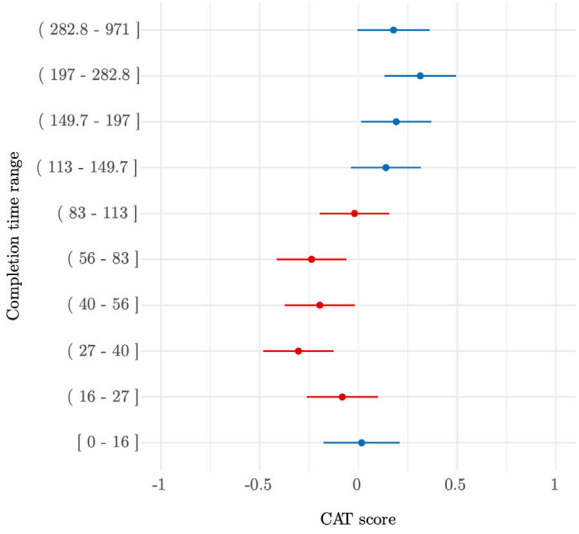



**Fig. 13. Task completion time and performance variations.** Each point represents the deviation of task completion time intervals from the average CAT score, with blue indicating scores above average and red below. Horizontal lines represent the estimates' confidence intervals.

Examining task completion times across age categories in Table 13, we found that older students do not necessarily complete tasks faster, contrary to expectations. Mean completion times generally increase with age, indicating a possible connection between age and task duration. However, this relationship is intricate, and interaction methods also play a crucial role. As shown in Fig. 3, older students use more advanced artefacts with higher autonomy, likely contributing to longer task resolution times. The 7–9 years age group, while employing interaction methods similar to those of the 3–6 years age group, accomplishes tasks at a relatively quicker pace. This implies a level of efficiency or adaptability within this age group, and the difference may be influenced by the rapid developmental changes occurring between these age groups.

In Table 14, the analysis of the times spent on specific interaction dimensions reveals some clear patterns. Users spend less time on the G interface, indicating lower preference or quicker navigation. On the contrary, more time is spent on the PF interface, suggesting higher preference or naturally longer interactions. The wider IQR indicates variability in user engagement with this interface. Comparing interfaces, users spend less time on gesture interfaces (G) compared to programming interfaces (P). Similarly, when relying on visual feedback, users spend less time with GF than with PF. This suggests that gesture-based interactions may be more efficient or intuitive, leading to quicker user engagement. Moreover, it appears users need more time when they have less autonomy, especially when relying on visual feedback. This could be due to the additional time required for students to continuously monitor and adjust their actions in response to visual feedback.

#### 4.4.7. Student performance dynamics in virtual and unplugged settings

To understand the factors influencing student performance in both unplugged and virtual settings, we combined the datasets to formulate the final model (M4), defined in (4), aimed at assessing how the domain impacts CAT scores, along with other contributing factors. The difference form M3 is the inclusion of the variable domain as a predictor of the CAT score.

$$\begin{aligned}\texttt{CAT\_SCORE} = &\beta_0 + \beta_1 \cdot \texttt{DOMAIN} + (\beta_2 + \beta_3 \cdot \texttt{SEX} + \epsilon_{\texttt{SEX}} | \texttt{SCHOOL}) + \\ &+ u_{\texttt{STUDENT}} + u_{\texttt{SESSION\_GRADE}} + u_{\texttt{SCHEMA}} + \epsilon.\end{aligned} \tag{4}$$

Variations in student performance across different groups, including individual students, session grades, schemas, and sex across schools, align with patterns identified in the model (M3) on individual datasets. Nevertheless, the model revealed that the domain effect on CAT scores lacked statistical significance ($p = 0.807$), thereby strengthening the coherence of these results across various settings. This underscores the robustness of the conclusion, highlighting the significance of domain-independent factors in shaping CAT scores.

The ANOVA results in Table 15 highlight that all factors – schema, sex within schools, session grade, and individual student traits – significantly influence CAT scores in both virtual and unplugged settings.

**Table 15**
**Type III analysis of variance (ANOVA) table with Satterthwaite's method on the full dataset.** Rows shaded in grey indicate statistically significant models.

| | AIC[a] | LRT[b] | $p$[c] |
|---|---|---|---|
| | 7461.0 | | |
| Schema | 7761.2 | 302.16 | <1e-15**** |
| Sex[d] | 7470.7 | 13.64 | 0.001** |
| Session-Grade | 7463.0 | 3.96 | 0.046* |
| Student | 8003.2 | 544.16 | <1e-15**** |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$

[a] Akaike Information Criterion for the model evaluated as $-2 \cdot (\text{logLik} - \text{npar})$. Smaller is better.

[b] LRT statistic; twice the difference in log-likelihood, which is asymptotically chi-square distributed.

[c] $p$ value $= Pr(> \chi^2)$

[d] Sex in (1 + Sex | School)

Comparing three models, M5 (reduced without the sex predictor), M6 (sex as a fixed effect), and M4 (sex as a random slope within

**Table 16**
**LRT to evaluate the global effect of Sex on the full dataset.** Comparison between the reduced model (M5) without the sex predictor, the model (M6) that considers sex a fixed effect, and the initial model (M4) that considers sex as a random slope within schools. Rows shaded in grey indicate statistically significant models.

| Model | AIC[a] | $\chi^2$ | $p$[b] |
|---|---|---|---|
| M5 | 7472.2 | | |
| M6 | 7473.6 | 0.631 | 0.4270 |
| M4 | 7462.5 | 13.063 | 0.0003*** |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$
[a] Akaike Information Criterion for the model evaluated as $-2 \cdot (\text{logLik} - \text{npar})$. Smaller is better.
[b] $p$ value $= Pr(> \chi^2)$

schools), we aimed to elucidate the role of sex in influencing CAT scores across diverse domains. The likelihood ratio test (LRT) results in Table 16 show that introducing sex as a fixed effect in M6 does not significantly improve the model compared to the baseline M5 ($p = 0.4270$). However, the inclusion of sex as a random slope within schools in M4 significantly enhances the model fit ($p = 1e-3^{***}$), emphasising that the impact of sex on CAT scores varies across different school environments. This underscores the importance of considering the interaction between sex and the school context when assessing its effect on educational outcomes.

## 5. Discussion and conclusion

In this section, we discuss the results to address our research questions. To do so, the discussion is structured into four subsections, each presenting the findings related to each specific research question. This is followed by the presentation of the limitations of the study, future works and recommendations.

### *5.1. What are the baseline competencies in AT in compulsory education, and how do they develop across school grades?*

#### *5.1.1. Contextual overview of findings*

Our analysis reveals a clear developmental trajectory in algorithmic thinking (AT) competencies with age across compulsory education. Younger students predominantly employ trial-and-error (T&E) strategies when encountering new concepts. As students advance in age and gain experience and maturity, they gradually shift towards more sophisticated problem-solving techniques and reduced reliance on T&E methods. This progression reflects a maturation in both problem-solving approaches and conceptual understanding.

The observed shift from T&E strategies to more refined problem-solving techniques suggests that as students develop, they become better equipped to handle complex algorithmic tasks. Younger students' preference for T&E can be seen as a necessary stage in their cognitive development, where iterative attempts and experience play a crucial role in their learning process. Over time, with increased experience and cognitive maturity, students can adopt more effective strategies that reflect deeper conceptual understanding and enhanced AT.

Our findings align with existing research highlighting a developmental progression in AT and problem-solving approaches with age (Del Olmo-Muñoz et al., 2020; El-Hamamsy, Bruno, Audrin, Chevalier, Avry, Zufferey, & Mondada, 2023; Kong & Lai, 2022; Román-González et al., 2017; Vlachogianni & Tselios, 2021). Similar to previous studies, our results confirm that younger students rely more on T&E approaches, especially when solutions are not immediately clear, and that as they mature, their problem-solving techniques become more sophisticated (Chevalier, Giang, Piatti, & Mondada, 2020; Kanaki & Kalogiannakis, 2022; Tónnsen, 2021). However, our study extends this understanding by detailing how the frequency of T&E behaviour impacts performance outcomes. While prior research has noted the shift from T&E to advanced strategies, our findings emphasise the nuanced role of T&E in both initial learning and subsequent algorithmic competence. While excessive T&E can initially impede performance, the process of iterative attempts can lead to improvement and adaptive learning. However, relying solely on T&E without integrating reflective thinking may hinder deeper conceptual understanding and the development of accurate algorithms (Chevalier et al., 2020; Shute, Sun, & Asbell-Clarke, 2017). This adds depth to understanding how iterative problem-solving can influence learning trajectories.

#### *5.1.2. Implications*

The developmental trajectory in AT competencies observed in our analysis underscores the importance of adapting instructional strategies to the cognitive maturity of students. The shift from T&E approaches to more sophisticated problem-solving techniques implies that educators should recognise and support the evolving nature of students' algorithmic skills. Younger students' reliance on T&E can be seen as an essential phase in their cognitive development, serving as a foundation for more advanced problem-solving methods. This progression highlights the need for educational frameworks that foster iterative learning while gradually introducing more complex algorithms and strategies. Understanding that T&E plays a dual role, both as a necessary step in learning and a potential barrier to advanced problem-solving, can help educators design curricula that balance exploration with reflective thinking to enhance students' overall algorithmic competence.

### *5.2. How do characteristics specific to the assessment instrument, such as different interaction modalities used in unplugged and digital instructional strategies, influence the development of AT skills in relation to sex, educational environment (e.g., school level and grade), and regional factors (e.g., the canton of the school)?*

#### *5.2.1. Contextual overview of findings*

*Impact of interaction modalities on AT skills development.* Our analysis reveals significant variations in how different interaction modalities impact AT skills across age groups. Younger students predominantly engage with simpler artefacts and use T&E strategies. As they grow older, they transition to more complex artefacts and show decreased reliance on T&E, which indicates a developmental progression towards more sophisticated problem-solving techniques. Interestingly, younger learners can also effectively engage with complex artefacts, demonstrating early development of advanced algorithmic skills. This finding suggests that students can benefit from exposure to complex artefacts and interactions even at a young age. This indicates the possibility that advanced cognitive skills can develop earlier than traditionally assumed.

Moreover, our results highlight that digital artefacts, due to their interactive and immersive nature, may be more effective than unplugged artefacts in fostering sophisticated AT skills, as observed with the CAT. The dynamic and engaging environment of virtual tools provides enhanced opportunities for algorithmic exploration and development, thereby supporting advanced cognitive growth in young learners. However, it is important to note that these findings are specific to this context and may not be applicable to all cases. Further research is necessary to confirm their broader relevance.

These findings contribute to a deeper theoretical understanding of how different interaction modalities influence cognitive development in AT. Our results are consistent with those obtained in the previous investigation on the unplugged CAT (Piatti et al., 2022) and support the literature indicating the rapid development of AT skills in preschool-aged children (Dietz, Landay, & Gweon, 2019; Nikolopoulou & Tsimperidis, 2023; Voronina et al., 2016; Vujičić, Jančec, & Mezak, 2021; Wahyuningsih, Nurjanah, Rasmani, Hafidah, Pudyaningtyas, & Syamsuddin, 2020). This alignment with prior studies underscores the notion that complex problem-solving abilities can emerge earlier than

previously thought (Georgiou & Angeli, 2021; Kanaki & Kalogiannakis, 2022; Sarama & Clements, 2009),

Our results extend these findings by showing that the effectiveness of virtual artefacts in enhancing AT skills is even more pronounced than previously documented for unplugged activities (Wohl, Porter, & Clinch, 2015). The interactive and stimulating nature of virtual environments provides a richer learning experience, which aligns with theories emphasising the role of immersive learning environments in cognitive development (Lui, Not, & Wong, 2023; Makransky & Petersen, 2021).

*Sex differences in educational contexts.* Our investigation into sex differences in computer science (CS) education did not reveal a global effect of sex on AT performance. However, it did uncover significant interactions between artefact type, sex, and age, as well as variations at the school level in shaping algorithmic complexity. Notably, we observed that in virtual environments, simpler artefacts generally impact algorithmic complexity less for males than for females, whereas unplugged environments show no significant sex effect. Age moderates these gender differences, with males aged 10 to 13 outperforming females in unplugged settings but lagging in virtual ones. Moreover, school performance data indicate variability in sex effects across different institutions. Some show higher performance for males, while others show better results for females.

These findings suggest that the impact of sex on AT performance is multifaceted and shaped by multiple factors, including artefact type, age, and school environment. This aligns with existing research that explores how sex differences impact performance outcomes in CS education (Ardito, Czerkawski, & Scollins, 2020; Kong & Lai, 2022; Mouza, Pan, Yang, & Pollock, 2020; Plante, de la Sablonniére, Aronson, & Théorêt, 2013; Sun, Hu, & Zhou, 2022). Notably, the literature underscores the importance of early exposure to CS education and effective teacher preparation in mitigating gender gaps and enhancing equity (El-Hamamsy et al., 2023). Gender differences from a young age also contribute to performance disparities, highlighting the need for targeted interventions and supportive educational environments (Master, Meltzoff, & Cheryan, 2021). Moreover, school-specific factors such as pedagogical methods, institutional culture, and student cohort dynamics play a significant role in these variations, highlighting the complex interplay between these elements in shaping student performance (Rachmatullah, Vandenberg, & Wiebe, 2022; Wang & Hejazi Moghadam, 2017). Research emphasises that the quality of instruction, classroom management, and local educational practices are crucial in shaping student outcomes (El-Hamamsy et al., 2023; Wang, Guo, & Degol, 2019). These insights are consistent with findings that highlight the importance of considering local contexts in educational policies and practices, as academic achievements can vary significantly across different cultures and regions (Wang et al., 2019).

*Diversity and individual differences.* The wide range of performances highlights the individual differences influenced by personal abilities, learning preferences, and external circumstances. This diversity underscores the necessity for creating equitable learning environments that accommodate various needs and learning styles. Addressing these differences is crucial for ensuring that every student has the opportunity to succeed and develop their AT skills fully.

The recognition of individual differences and their impact on learning outcomes supports the growing body of literature advocating for personalised and adaptive educational approaches. This perspective aligns with research suggesting that personalised instruction tailored to individual needs and characteristics can enhance learning experiences (Desmarais & Baker, 2011; Hooshyar, Ahmad, Yousefi, Fathi, Horng, & Lim, 2016; Millán, Pérez-de-la Cruz, & Suárez, 2000; Mousavinasab, Zarifsanaiey, Niaka Kalhori, Rakhshan, Keikha, & Ghaz. Saeedi, 2018; Soofi & Uddin, 2019; Vomlel, 2004). By adapting educational practices to address diverse learning preferences and abilities, educators can create more inclusive and supportive environments that foster success for all students.

#### 5.2.2. Implications

The findings from our analysis underscore the significant impact that different interaction modalities and factors such as sex, educational environment (including school level and grade), and regional context have on the development of AT skills.

Our results suggest that virtual artefacts can be more effective than unplugged ones in fostering sophisticated AT skills. This implies that educators and policymakers might consider integrating more interactive and immersive technologies into the curriculum to enhance cognitive development. However, it is important to balance this integration with considerations about screen time and its implications for young learners. For instance, while digital platforms can enhance learning by providing early exposure to computational concepts, excessive screen time has been linked to several potential issues. Excessive screen time can negatively affect cognitive development, executive functioning, and social-emotional skills, leading to issues like reduced academic performance, impaired language development, and increased risks of obesity and mental health problems (Muppalla, Vuppalapati, Redd. Pulliahgaru, & Sreenivasulu, 2023; Ponti, 2023; Swider-Cios, Vermeij, & Sitskoorn, 2023). To mitigate these risks, it is crucial to set reasonable limits on screen time and promote a balanced approach. The benefits of digital tools in early education must be weighed against these potential risks. By integrating interactive and immersive technologies thoughtfully, and balancing screen time with other developmental activities, we can support young learners' AT skills while promoting their overall well-being and healthy development.

Additionally, the observed variability in performance based on sex, age, and school environment highlights the need for nuanced educational strategies. Addressing these variations is crucial for mitigating biases and ensuring equitable access to resources. Creating inclusive and adaptive learning environments that cater to diverse needs and learning styles is key to supporting all students effectively.

### 5.3. Future directions and recommendations

To foster equitable learning environments, several key strategies should be implemented. First, curricula should integrate both exploratory and reflective learning opportunities. Curricula and learning experiences should be designed to accommodate diverse learning styles and preferences. Offering a variety of instructional materials and methods, such as virtual and unplugged activities, ensures that all students have access to resources that align with their individual learning needs. This involves developing age-appropriate instructional strategies that start with T&E methods and gradually introduce more structured problem-solving techniques. For younger students, the focus should be on iterative attempts and experiential learning, with more formal algorithmic concepts introduced as their cognitive abilities mature. Alongside T&E methods, incorporating activities that promote reflective thinking can help students analyse their iterative attempts, identify patterns, and evaluate alternative strategies, bridging the gap between basic problem-solving and more advanced techniques.

Instructional strategies must also be differentiated to accommodate variations in sex, age, and educational contexts, addressing specific needs and minimising biases. Adapting educational practices to local contexts and continuously assessing their effectiveness can help create more equitable learning environments. Additionally, future research should investigate how students' self-perception, interest, and motivation impact their performance in AT. Existing literature has established a strong link between high student engagement and a positive perception of the learning environment with increased academic success (Bellino & Herskovic, 2023; El-Hamamsy et al., 2023; Hinckle, Rachmatullah, Mott, Boyer, Lester, & Wiebe, 2020; Olivier, Archambault, D. Clercq, & Galand, 2018; Rachmatullah et al., 2022; Sun et al., 2022; Tai, Ryoo, Skeeles-Worley, Dabney, Almarode, & Maltese, 2022). Understanding these factors can provide insights into

how students' attitudes and internal perceptions influence their engagement and achievement, further informing tailored instructional strategies (Beyer, 2014; Guran, Cojocar, & Turian, 2020; Kong, Chiu, & Lai, 2018; Master et al., 2021; Olivier et al., 2018; Sevin & Decamp, 2016; Wang, Dai, & Mathis, 2022).

Educators play a crucial role in this process and should receive training on the developmental stages of AT and effective instructional practices tailored to each stage. This professional development should also include strategies for adapting teaching methods based on students' backgrounds to create inclusive learning environments. Additionally, formative assessments should monitor students' progress and adapt teaching methods accordingly, ensuring that instructional practices address individual learning needs and support effective transitions from T&E to advanced techniques.

Virtual artefacts and interactive tools are vital in providing dynamic and engaging learning experiences, particularly for younger students. These technologies can significantly enhance cognitive growth and algorithmic skills.

Lastly, while our study did not specifically explore the practical implementation of personalised instruction based on AT skills in real-world classrooms, we can propose methods for integrating these findings into everyday educational settings. Recognising that a uniform approach may not be effective for all students highlights the importance of tailoring instruction to individual needs and characteristics. Intelligent Tutoring and Assessment Systems (ITAS) can play a key role here (Rodriguez-Barrios, Melendez-Armenta, Garcia-Aburto, Lavoignet-Ruiz, Sandoval-Herazo, Molina-Navarro, & Morales-Rosales, 2021; Wu, 2019; Xing et al., 2020). These adaptive systems that suggest tasks aligned with students' proficiency levels and provide automatic tutoring mechanisms can further enhance learning experiences (Millán et al., 2000; Soofi & Uddin, 2019; Vomlel, 2004). By offering recommendations for different artefacts or visual feedback, these systems support individualised instruction and contribute to a more equitable educational system, ensuring that all students have the opportunity to succeed (Desmarais & Baker, 2011; Hooshyar et al., 2016; Mousavinasab et al., 2018).

In this context, an attempt was made to build an Intelligent Assessment System (IAS) for the unplugged CAT using Bayesian networks, without incorporating tutoring features. Bayesian networks represent a promising future direction for assessment due to their ability to offer detailed, probabilistic evaluations of students' skills, as opposed to current methods that provide a single score per student-task (Antonucci, Mangili, Bonesana, & Adorni, 2022), allowing for a comprehensive learner model based on posterior probabilities. This approach, validated in two studies (Adorni, Mangili, Piatti, Bonesana, & Antonucci, 2023a; Mangili, Adorni, Piatti, Bonesana, & Antonucci, 2022), can be adapted to the virtual CAT environment, offering a more nuanced way of assessing students. Integrating tutoring capabilities into this Bayesian network-based IAS could evolve it into a fully-fledged ITAS, delivering real-time, adaptive support for students. Future research should focus on refining these models, integrating them into virtual CAT assessment tools, and incorporating tutoring functionalities to enhance their effectiveness and scalability in diverse educational contexts.

### 5.4. Limitations

In this section, we address the study's potential limitations, discussing how we have attempted to mitigate some of these challenges and acknowledging the limitations that remain unresolved.

One significant limitation is the exclusive use of the CAT as the sole assessment instrument. This approach restricts the ability to benchmark the CAT's effectiveness against alternative measures of AT without comparing it to other validated assessment tools. Future research should incorporate additional AT assessment tools to provide a broader perspective and validate the results obtained with the CAT to enhance the robustness of the findings. Comparing outcomes from different assessment methods could offer a more comprehensive understanding of AT development.

A possible limitation of the study is the variability in implementing instructional strategies, particularly the differences in administering tutorials across various linguistic regions within Switzerland. This inconsistency may have influenced the effectiveness and comparability of the outcomes. For example, differences in how tutorials were administered by different administrators in the Italian-speaking and German-speaking regions could have introduced variations in the delivery of instructional materials and the administration of the CAT, potentially impacting the study's results. To address this issue, we have proactively redesigned the training modules to standardise the administration method. We developed an in-app video tutorial system to allow users to navigate the platform independently, thus reducing potential biases from varying researcher explanations. While this enhancement has been integrated into the system, it has not yet been tested in practice. Future research will benefit from evaluating this standardised approach to ensure a consistent experience for all participants and improve the reliability of the results.

Another possible limitation of the study is the potential lack of socio-economic diversity within the sample, which may affect the unreliability of the findings. Although we aimed to include a diverse sample by balancing factors such as sex, age, educational environment (e.g., school level and grade), and regional factors (e.g., the canton of the school), we did not specifically investigate socio-economic factors like parental income or education. The non-random selection process, where schools and classes were recruited through agreements with school directors and teachers, might have introduced selection bias. Additionally, the limited sample size constrains our ability to generalise the results to a broader population.

Furthermore, the study did not account for the extent of students' prior digital education, which could significantly influence their performance on the assessment. Previous exposure to digital learning tools and environments may affect students' familiarity with the technology used in the study, potentially impacting their ability to engage with and benefit from the instructional strategies tested. Without considering this variable, the findings may not fully capture the interplay between prior digital experience and AT development. This oversight could limit the understanding of how previous digital education affects students' performance and the effectiveness of the proposed instructional methods. Future research should address this by collecting data on students' prior digital education and exploring its impact on learning outcomes to provide a more comprehensive view of how previous experiences shape AT development.

Despite these limitations, we believe the findings are relevant beyond the Swiss context. The diverse nature of the sample, which included students from various geographic and linguistic regions within Switzerland, and the absence of specific characteristics that would make these classes unusually different from those in highly educated countries suggest that the results may apply to other educational environments and cultural settings. Nevertheless, future research should continue exploring these findings' applicability in different contexts to confirm their broader relevance.

Finally, a potential limitation of the study is related to technological access and resource availability. Although we ensured that all participants had access to the necessary technology for this study, real-world implementation might face constraints due to varying levels of technological access and resource availability in different educational settings. Differences in the quality of technology and resources could influence the effectiveness of virtual artefacts and interactive tools beyond the controlled environment of the study. To address this, future research should investigate how varying levels of technological support and resource constraints impact the implementation and effectiveness of educational interventions, ensuring that findings apply to a broader range of educational contexts.

## Software availability

The software components used in this study are open-source – virtual CAT platform (Adorni, Piatti, & Karpenko, 2023b), virtual CAT programming language interpreter (Adorni & Karpenko, 2023c), virtual CAT data infrastructure (Adorni & Karpenko, 2023d) –, as is the code to reproduce the results (Adorni, 2024b).

## CRediT authorship contribution statement

**Giorgia Adorni:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Igor Artico:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Alberto Piatti:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Elia Lutz:** Writing – review & editing, Investigation. **Luca Maria Gambardella:** Writing – review & editing, Supervision, Funding acquisition. **Lucio Negrini:** Writing – review & editing. **Francesco Mondada:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Dorit Assaf:** Writing – review & editing, Funding acquisition.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Giorgia Adorni reports financial support was provided by Swiss National Science Foundation (SNSF) under the National Research Programme on Digital Transformation NRP-77 (407740_187246). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The data supporting this research is open-source and includes the protocol template for administering the unplugged CAT assessment (Piatti & Adorni, 2024) and the dataset from the virtual CAT assessment (Adorni, 2024a).

## Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT and Grammarly to enhance language and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

## Acknowledgments

We would like to express our gratitude to Francesca Mangili, Kunal Massé and Jérôme Guillaume Brender for their valuable feedback and support during the application testing and refinement process.

## Appendix A

See Figs. A.1–A.3.

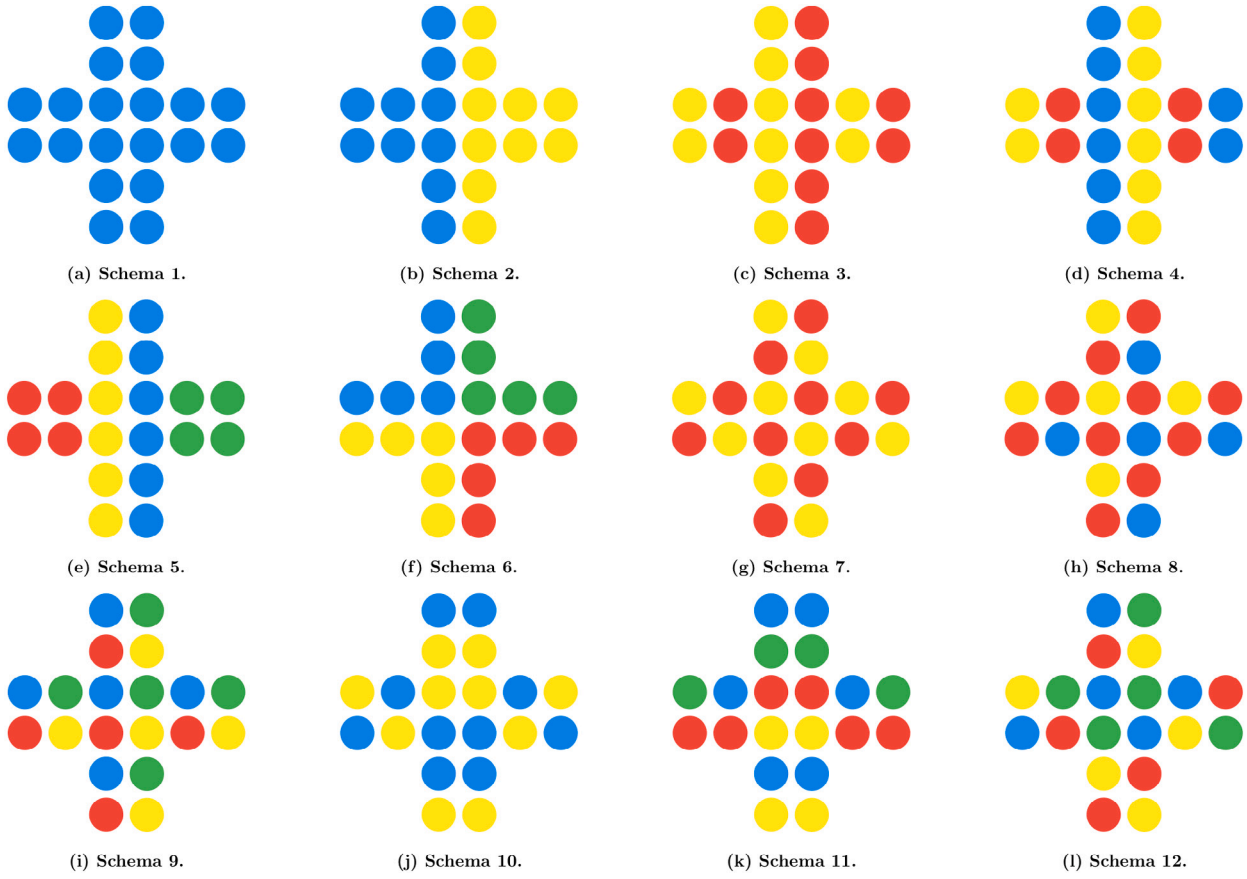



**Fig. A.1. Sequence of CAT schemas.** The 12 schemas proposed in the task, named from Schema 1 to Schema 12, are characterised by unique visual regularities and complexities, varying in elements such as colours, symmetries, alternations and other distinctive features.
*Source:* Adapted from Piatti et al. (2022).

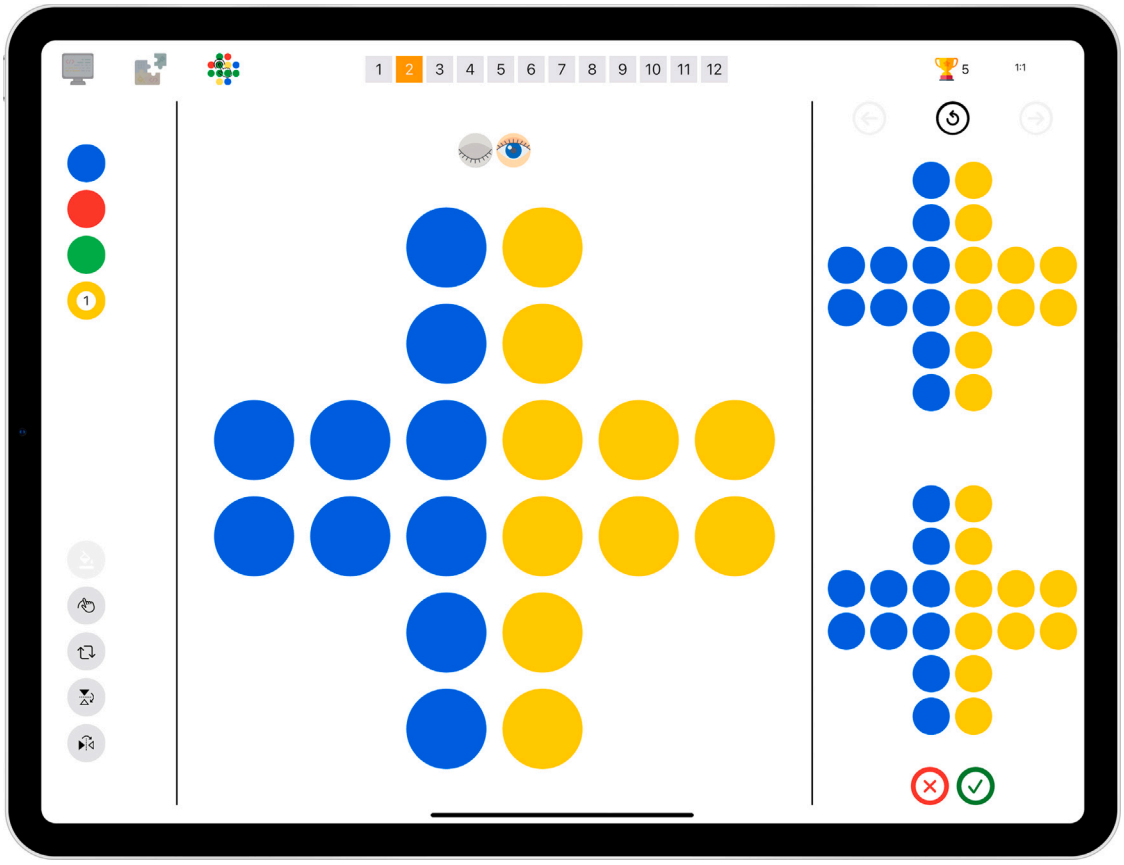



**Fig. A.2. Example of usage of the CAT-VPI with textual commands.**

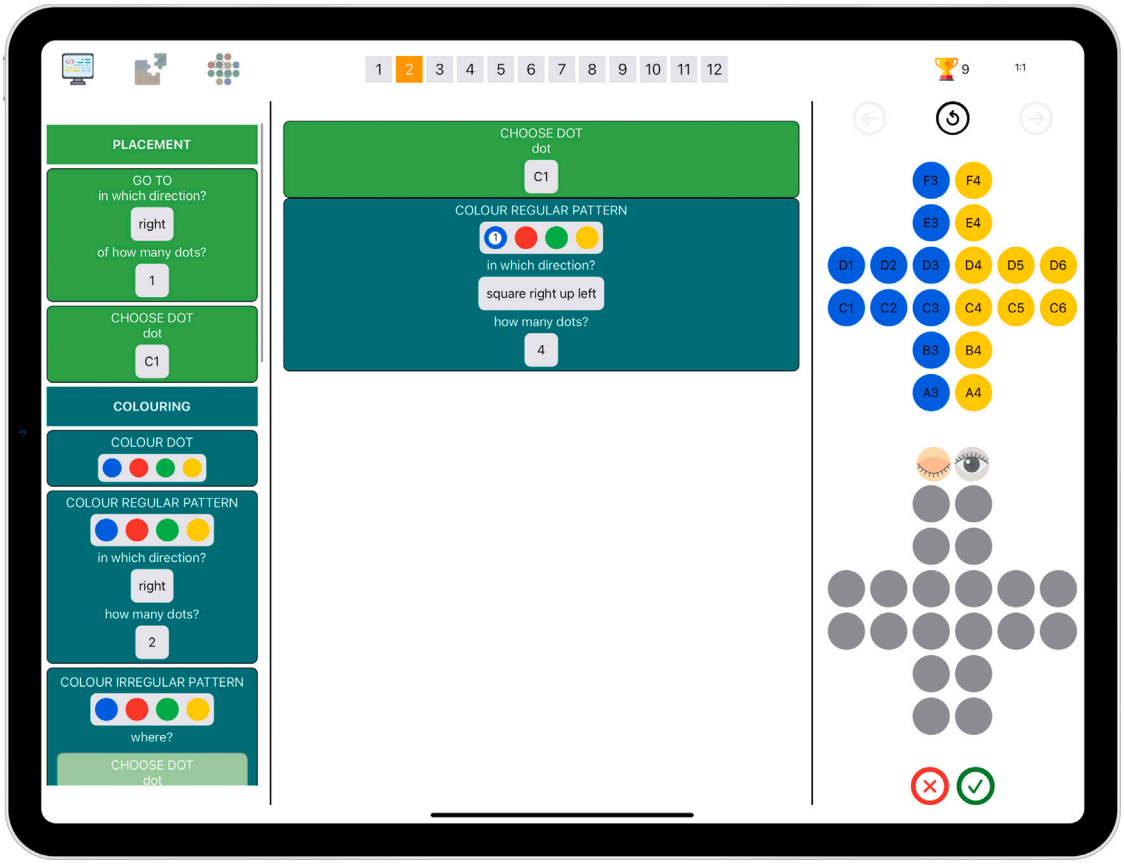



**Fig. A.3. Example of usage of the CAT-GI.**

## References


Adorni, G. (2024a). Dataset for algorithmic thinking skills assessment: Results from the virtual CAT large-scale study in Swiss compulsory education. http://dx.doi.org/10.5281/zenodo.10912339.

Adorni, G (2024b). Virtual CAT Algorithmic Thinking Assessment: Data Analysis Procedures. http://dx.doi.org/10.5281/zenodo.12805318.

Adorni, G., & Karpenko, V. (2023c). virtual CAT programming language interpreter. http://dx.doi.org/10.5281/zenodo.10016535.

Adorni, G., & Karpenko, V. (2023d). virtual CAT data infrastructure. http://dx.doi.org/10.5281/zenodo.10015011.

Adorni, G., Mangili, F., Piatti, A., Bonesana, C., & Antonucci, A. (2023a). Rubric-based Learner modelling via noisy gates Bayesian networks for computational thinking skills assessment. *Journal of Communications Software and Systems*, *19*, 52–64. http://dx.doi.org/10.24138/jcomss-2022-0169.

Adorni, G., & Piatti, A. (2024c). The virtual CAT: A tool for algorithmic thinking assessment in Swiss compulsory education. http://dx.doi.org/10.48550/arXiv.2408.01263, arXiv:2408.01263.

Adorni, G., Piatti, A., Bumbacher, E., Negrini, L., Mondada, F., Assaf, D., et al. (2024d). A theoretical framework for the design and analysis of computational thinking problems in education. http://dx.doi.org/10.48550/arXiv.2403.19475.

Adorni, G., Piatti, S., & Karpenko, V. (2023b). virtual CAT: An app for algorithmic thinking assessment within Swiss compulsory education. http://dx.doi.org/10.5281/zenodo.10027851.

Adorni, G., Piatti, S., & Karpenko, V. (2024e). Virtual CAT: A multi-interface educational platform for algorithmic thinking assessment. *SoftwareX*, *27*, 101737. http://dx.doi.org/10.1016/j.softx.2024.101737.

Aebi-Müller, R. E., Blatter, I., Brigger, J., Constable, E. C., Eglin, N., Hoffmeyer, P., et al. (2021). Code of conduct for scientific integrity. http://dx.doi.org/10.5281/zenodo.4707560.

Antonucci, A., Mangili, F., Bonesana, C., & Adorni, G. (2022). Intelligent Tutoring Systems by Bayesian Nets with Noisy Gates. *The International FLAIRS Conference Proceedings*, *35*, http://dx.doi.org/10.32473/flairs.v35i.130692.

Ardito, G., Czerkawski, B., & Scollins, L. (2020). Learning computational thinking together: Effects of gender differences in collaborative middle school robotics program. *TechTrends*, *64*, 373–387. http://dx.doi.org/10.1007/s11528-019-00461-8.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, *160*, 268–282. http://dx.doi.org/10.1098/rspa.1937.0109.

Bell, T., Alexander, J., Freeman, I., & Grimley, M. (2009). Computer science unplugged: School students doing real computing without computers. *The New Zealand Journal of Applied Computing and Information Technology*, *13*(1), 20–29, https://purehost.bath.ac.uk/ws/portalfiles/portal/214932627/NZJACIT_Unplugged.pdf.

Bell, T., & Vahrenhold, J. (2018). Cs unplugged–how is it used, and does it work? In *Lecture notes in computer science* (pp. 497–521). Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-98355-4_29.

Bellino, A., & Herskovic, V. (2023). Protobject as a tool for teaching computational thinking to designers: student perceptions on usability. In *Proceedings of the 15th biannual conference of the Italian SIGCHI chapter, CHItaly 2023*. ACM, http://dx.doi.org/10.1145/3605390.3605401.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*, 289–300. http://dx.doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Bers, M. U., Strawhacker, A., & Sullivan, A. (2022). The state of the field of computational thinking in early childhood education. http://dx.doi.org/10.1787/3354387a-en, OECD Education Working Papers 274 OECD Publishing.

Beyer, S. (2014). Why are women underrepresented in computer science? gender differences in stereotypes, self-efficacy, values, and interests and predictors of future cs course-taking and grades. *Computer Science Education*, *24*, 153–192. http://dx.doi.org/10.1080/08993408.2014.963363.

Bland, J. M., & Altman, D. G. (1995). Statistics notes: Multiple significance tests: The bonferroni method. *BMJ*, *310*, 170. http://dx.doi.org/10.1136/bmj.310.6973.170.

Bocconi, S., Chioccariello, A., Kampylis, P., Dagien, V., Wastiau, P., Engelhardt, K., et al. (2022). *Reviewing computational thinking in compulsory education*: *Technical report*, Joint Research Centre (Seville site, http://dx.doi.org/10.2760/126955.

Brackmann, C. P., Román-González, M., Robles, G., Moreno-León, J., Casali, A., & Barone, D. (2017). Development of Computational Thinking Skills through Unplugged Activities in Primary School. In *Proceedings of the 12th workshop on primary and secondary computing education* (pp. 65–72). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3137065.3137069.

Campbell-Barr, V., Lavelle, M., & Wickett, K. (2012). Exploring alternative approaches to child outcome assessments in children' centres. *Early Child Development and Care*, *182*, 859–874. http://dx.doi.org/10.1080/03004430.2011.590937.

Chambers, J., Hastie, T., & Pregibon, D. (1990). Statistical models in S. In *Compstat* (pp. 317–321). Physica-Verlag HD, http://dx.doi.org/10.1007/978-3-642-50096-1_48.

Chevalier, M., Giang, C., Piatti, A., & Mondada, F. (2020). Fostering computational thinking through educational robotics: A model for Creative Computational Problem Solving (CCPS). *International Journal of STEM Education*, *39*, http://dx.doi.org/10.1186/s40594-020-00238-z.

Cochran, W. G. (1954). Some methods for strengthening the common $\chi^2$ tests. *Biometrics*, *10*(417), http://dx.doi.org/10.2307/3001616.

Cox, D., & Hinkley, D. (1979). *Theoretical statistics*. D. Chapman and Hall/CRC, http://dx.doi.org/10.1201/b14832.

Csernoch, M., Biró, P., Máth, J., & Abari, K. (2015). Testing algorithmic skills in traditional and non-traditional programming environments. *Informatics in Education*, *14*, 175–197. http://dx.doi.org/10.15388/infedu.2015.11.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511802843.

Del Olmo-Muñoz, J., Cózar-Gutiérrez, R., & González-Calero, J. A. (2020). Computational thinking through unplugged activities in early years of Primary Education. *Computers & Education*, *150*, Article 103832. http://dx.doi.org/10.1016/j.compedu.2020.103832.

Desmarais, M. C., & Baker, R. S. J. d. (2011). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, *22*, 9–38. http://dx.doi.org/10.1007/s11257-011-9106-8.

Dietz, G., Landay, J. A., & Gweon, H. (2019). Building blocks of computational thinking: Young children's developing capacities for problem decomposition. In *Annual meeting of the cognitive science society*. https://api.semanticscholar.org/CorpusID:198232922.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*, 52–64. http://dx.doi.org/10.1080/01621459.1961.10482090.

El-Hamamsy, L., Bruno, B., Audrin, C., Chevalier, M., Avry, S., Zufferey, J. D., et al. (2023). How are primary school computer science curricular reforms contributing to equity? Impact on student learning, perception of the discipline, and gender gaps. *International Journal of STEM Education*, *10*, http://dx.doi.org/10.1186/s40594-023-00438-3.

El-Hamamsy, L., Zapata-Cáceres, M., Barroso, E. M., Mondada, F., Zufferey, J. D., & Bruno, B. (2022). The competent computational thinking test: Development and validation of an unplugged computational thinking test for upper primary school. *Journal of Educational Computing Research*, *60*, 1818–1866. http://dx.doi.org/10.1177/07356331221081753.

Ezeamuzie, N. O., & Leung, J. W. (2021). Computational thinking through an empirical lens: A systematic review of literature. *Journal of Educational Computing Research*, *60*, 481–511. http://dx.doi.org/10.1177/07356331211033158.

Fisk, P. R., & Weisberg, S. (1982). Applied linear regression. *Journal of the Royal Statistical Society. Series A (General)*, *145*(146), http://dx.doi.org/10.2307/2981445.

Georgiou, K., & Angeli, C. (2021). Developing computational thinking in early childhood education: A focus on algorithmic thinking and the role of cognitive differences and scaffolding. In D. Ifenthaler, D. G. Sampson, & P. Isaías (Eds.), *Balancing the tension between digital technologies and learning sciences* (pp. 33–49). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-65657-7_3.

Grover, S. (2017). Assessing algorithmic and computational thinking in k-12: Lessons from a middle school classroom. In P. J. Rich, & C. B. Hodges (Eds.), *Emerging research, practice, and policy on computational thinking* (pp. 269–288). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-52691-1_17.

Guran, A.-M., Cojocar, G.-S., & Turian, A. (2020). Towards preschoolers' automatic satisfaction assessment. An experience report. In *2020 IEEE 14th international symposium on applied computational intelligence and informatics*. IEEE, http://dx.doi.org/10.1109/saci49304.2020.9118824.

Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The elements of statistical learning*. New York: Springer, http://dx.doi.org/10.1007/978-0-387-21606-5.

Hinckle, M., Rachmatullah, A., Mott, B., Boyer, K. E., Lester, J., & Wiebe, E. (2020). The relationship of gender, experiential, and psychological factors to achievement in computer science. In *Proceedings of the 2020 ACM conference on innovation and technology in computer science education*. ACM, http://dx.doi.org/10.1145/3341525.3387403.

Hooshyar, D., Ahmad, R. B., Yousefi, M., Fathi, M., Horng, S.-J., & Lim, H. (2016). Sits: A solution-based intelligent tutoring system for students' acquisition of problem-solving skills in computer programming. *Innovations in Education and Teaching International*, *55*, 325–335. http://dx.doi.org/10.1080/14703297.2016.1189346.

Hox, J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge, http://dx.doi.org/10.4324/9781315650982.

Hsu, T.-C., Chang, S.-C., & Hung, Y.-T. (2018). How to learn and how to teach computational thinking: Suggestions based on a review of the literature. *Computers & Education*, *126*, 296–310. http://dx.doi.org/10.1016/j.compedu.2018.07.004.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer, http://dx.doi.org/10.1007/978-1-4614-7138-7.

Jiang, S., & Wong, G. K. W. (2022). Exploring age and gender differences of computational thinkers in primary school: A developmental perspective. *Journal of Computer Assisted Learning*, *38*, 60–75. http://dx.doi.org/10.1111/jcal.12591.

Kalelioglu, F., Gulbahar, Y., & Kukul, V. (2016). A Framework for computational thinking based on a systematic research review. *Baltic Journal of Modern Computing*, *4*, 583–596, https://api.semanticscholar.org/CorpusID:26908185.

Kanaki, K., & Kalogiannakis, M. (2022). Assessing algorithmic thinking skills in relation to age in early childhood stem education. *Education Sciences*, *12*, http://dx.doi.org/10.3390/educsci12060380.

Keith, P. K., Sullivan, F. R., & Pham, D. (2019). Roles, collaboration, and the development of computational thinking in a robotics learning environment. In *Computational thinking education* (pp. 223–245). Singapore: Springer, http://dx.doi.org/10.1007/978-981-13-6528-7_13.

Kong, S.-C., Chiu, M. M., & Lai, M. (2018). A study of primary school students' interest, collaboration attitude, and programming empowerment in computational thinking education. *Computers & Education*, *127*, 178–189. http://dx.doi.org/10.1016/j.compedu.2018.08.026.

Kong, S.-C., & Lai, M. (2022). Validating a computational thinking concepts test for primary education using item response theory: An analysis of students' responses. *Computers & Education*, *187*, Article 104562. http://dx.doi.org/10.1016/j.compedu.2022.104562.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, http://dx.doi.org/10.18637/jss.v082.i13.

Lenth, R. V. (2023). Emmeans: Estimated marginal means, aka least-squares means. R package version 1.9.0; https://CRAN.R-project.org/package=emmeans.

Lui, A. L. C., Not, C., & Wong, G. K. W. (2023). Theory-based learning design with immersive virtual reality in science education: A systematic review. *Journal of Science Education and Technology*, *32*, 390–432. http://dx.doi.org/10.1007/s10956-023-10035-2.

Makransky, G., & Petersen, G. B. (2021). The cognitive affective model of immersive learning (camil): A theoretical research-based model of learning in immersive virtual reality. *Educational Psychology Review*, *33*, 937–958. http://dx.doi.org/10.1007/s10648-020-09586-2.

Mangili, F., Adorni, G., Piatti, A., Bonesana, C., & Antonucci, A. (2022). Modelling Assessment Rubrics through Bayesian Networks: A Pragmatic Approach. In *2022 international conference on software, telecommunications and computer networks*. IEEE, http://dx.doi.org/10.23919/softcom55329.2022.9911432.

Martin, N., & Maes, H. (1979). *Multivariate analysis*. London, UK: Academic, https://ibg.colorado.edu/workshop2008/cdrom/ScriptsI/maes/Multivariate/Multivariate-mac.pdf.

Master, A., Meltzoff, A. N., & Cheryan, S. (2021). Gender stereotypes about interests start early and cause gender disparities in computer science and engineering. *Proceedings of the National Academy of Sciences*, *118*, http://dx.doi.org/10.1073/pnas.2100030118.

McCormick, K. I., & Hall, J. A. (2022). Computational thinking learning experiences, outcomes, and research in preschool settings: A scoping review of literature. *Education and Information Technologies*, 1–36. http://dx.doi.org/10.1007/s10639-021-10765-z.

Millán, E., Pérez-de-la Cruz, J. L., & Suárez, E. (2000). Adaptive Bayesian networks for multilevel student modelling. In *Lecture Notes in Computer Science*, (pp. 534–543). Berlin Heidelberg: Springer, http://dx.doi.org/10.1007/3-540-45108-0_57.

Moore, D. S., & McCabe, G. P. (1989). *Introduction to the practice of statistics*. New York, NY, US: W.H. Freeman, https://books.google.ch/books?id=pGBNhajABlUC.

Mousavinasab, E., Zarifsanaiey, N. R., Niaka Kalhori, S., Rakhshan, M., Keikha, L., & Ghaz. Saeedi, M. (2018). Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, *29*, 142–163. http://dx.doi.org/10.1080/10494820.2018.1558257.

Mouza, C., Pan, Y.-C., Yang, H., & Pollock, L. (2020). A multiyear investigation of student computational thinking concepts, practices, and perspectives in an after-school computing program. *Journal of Educational Computing Research*, *58*, 1029–1056. http://dx.doi.org/10.1177/0735633120905605.

Muppalla, S. K., Vuppalapati, S., Redd. Pulliahgaru, A., & Sreenivasulu, H. (2023). Effects of excessive screen time on child development: An updated review and strategies for management. *Cureus*, http://dx.doi.org/10.7759/cureus.40608.

Newcombe, R. G. (1998a). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, *17*, 873–890. http://dx.doi.org/10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I.

Newcombe, R. G. (1998b). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine*, *17*, 857–872. http://dx.doi.org/10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E.

Nikolopoulou, K., & Tsimperidis, I. (2023). Stem education in early primary years: Teachers' views and confidence. *Journal of Digital Educational Technology*, *3*, ep2302. http://dx.doi.org/10.30935/jdet/12971.

Olivier, E., Archambault, I., D. Clercq, M., & Galand, B. (2018). Student self-efficacy, classroom engagement, and academic achievement: Comparing three theoretical frameworks. *Journal of Youth and Adolescence*, *48*, 326–340. http://dx.doi.org/10.1007/s10964-018-0952-0.

Oyelere, S. S., Agbo, F. J., & Sanusi, I. T. (2022). Developing a pedagogical evaluation framework for computational thinking supporting technologies and tools. *Frontiers in Education*, *7*, http://dx.doi.org/10.3389/feduc.2022.957739.

Perneger, T. V. (1998). What's wrong with bonferroni adjustments. *BMJ*, *316*, 1236–1238. http://dx.doi.org/10.1136/bmj.316.7139.1236.

Petousi, V., & Sifaki, E. (2020). Contextualising harm in the framework of research misconduct, findings from discourse analysis of scientific publications. *International Journal of Sustainable Development*, *23*(149), http://dx.doi.org/10.1504/ijsd.2020.115206.

Piaget, J. (1964). Part I: Cognitive development in children: Piaget development and learning. *Journal of Research in Science Teaching*, *2*, 176–186. http://dx.doi.org/10.1002/tea.3660020306.

Piaget, J., Cook, M., et al. (1952). *The origins of intelligence in children*. W W Norton & Co, http://dx.doi.org/10.1037/11494-000.

Piaget, J., & Mussen, P. (1983). *History, theory, and methods*, *Handbook of Child Psychology*. https://books.google.ch/books?id=xe1GAAAAMAAJ.

Piatti, A., & Adorni, G. (2024). Unplugged Cross Array Task (CAT) Assessment: Supplementary Documentation and Experimental Protocol. http://dx.doi.org/10.5281/zenodo.12806226.

Piatti, A., Adorni, G., El-Hamamsy, L., Negrini, L., Assaf, D., Gambardella, L., et al. (2022). The CT-cube: A framework for the design and the assessment of computational thinking activities. *Computers in Human Behavior Reports*, *5*, Article 100166. http://dx.doi.org/10.1016/j.chbr.2021.100166.

Pilotti, M., Nazeeruddin, E., Mohammad, N., Daqqa, I., Abdelsalam, H., & Abdullah, M. M. (2022). Is initial performance in a course informative? machine learning algorithms as aids for the early detection of at-risk students. *Electronics*, *11*(2057), http://dx.doi.org/10.3390/electronics11132057.

Plante, I., de la Sablonniére, R., Aronson, J. M., & Théorêt, M. (2013). Gender stereotype endorsement and achievement-related outcomes: The role of competence beliefs and task values. *Contemporary Educational Psychology*, *38*, 225–235. http://dx.doi.org/10.1016/j.cedpsych.2013.03.004.

Ponti, M. (2023). Screen time and preschool children: Promoting health and development in a digital world. *Paediatrics & Child Health*, *28*, 184–192. http://dx.doi.org/10.1093/pch/pxac125.

Qian, Y., & Lehman, J. D. (2018). Using technology to support teaching computer science: A study with middle school students. *Eurasia Journal of Mathematics Science and Technology Education*, *14*, http://dx.doi.org/10.29333/ejmste/94227.

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.

Rachmatullah, A., Vandenberg, J., & Wiebe, E. (2022). Toward more generalizable CS and CT instruments: Examining the interaction of country and gender at the middle grades level. In *Proceedings of the 27th ACM conference on on innovation and technology in computer science education: vol. 1*, (pp. 179–185). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3502718.3524790.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods: vol. 1*, sage, https://books.google.ch/books?id=uyCV0CNGDLQC.

Relkin, E., de Ruiter, L., & Bers, M. U. (2020). TechCheck: Development and validation of an unplugged assessment of computational thinking in early childhood education. *Journal of Science Education and Technology*, *29*, 482–498. http://dx.doi.org/10.1007/s10956-020-09831-x.

Rodriguez-Barrios, E. U., Melendez-Armenta, R. A., Garcia-Aburto, S. G., Lavoignet-Ruiz, M., Sandoval-Herazo, L. C., Molina-Navarro, A., et al. (2021). Bayesian approach to analyze reading comprehension: A case study in elementary school children in Mexico. *Sustainability*, *13*(4285), http://dx.doi.org/10.3390/su13084285.

Román-González, M., Pérez-González, J.-C., & Jiménez-Fernández, C. (2017). Which cognitive abilities underlie computational thinking? criterion validity of the computational thinking test. *Computers in Human Behavior*, *72*, 678–691. http://dx.doi.org/10.1016/j.chb.2016.08.047.

Romero, M., Lepage, A., & Lille, B. (2017). Computational thinking development through creative programming in higher education. *International Journal of Educational Technology in Higher Education*, *14*, 1–15. http://dx.doi.org/10.1186/s41239-017-0080-z.

Sarama, J., & Clements, D. H. (2009). *Early Childhood Mathematics Education Research: Learning Trajectories for Young Children*. Routledge, http://dx.doi.org/10.4324/9780203883785.

Scherer, R., Siddiq, F., & Viveros, B. S. (2019). The cognitive benefits of learning computer programming: A meta-analysis of transfer effects. *Journal of Educational Psychology*, *111*, 764–792. http://dx.doi.org/10.1037/edu0000314.

Seber, G. A. F. (1984). *Multivariate observations*. Wiley, http://dx.doi.org/10.1002/9780470316641.

Sedgwick, P. (2014). Multiple hypothesis testing and Bonferroni's correction. *BMJ*, *349*, g6284. http://dx.doi.org/10.1136/bmj.g6284.

Sevin, R., & Decamp, W. (2016). From playing to programming: The effect of video game play on confidence with computers and an interest in computer science. *Sociological Research Online*, *21*, 14–23. http://dx.doi.org/10.5153/sro.4082.

Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review*, *22*, 142–158. http://dx.doi.org/10.1016/j.edurev.2017.09.003.

Silvey, S. (2017). *Statistical inference*. Routledge, http://dx.doi.org/10.1201/9780203738641.

Simmering, V. R., Ou, L., & Bolsinova, M. (2019). What technology can and cannot do to support assessment of non-cognitive skills. *Frontiers in Psychology*, *10*, http://dx.doi.org/10.3389/fpsyg.2019.02168.

SNSF (2021). Open science. https://www.snf.ch/en/dah3uC2QX95tfPNd/topic/open-science.

Soofi, A. A., & Uddin, M. (2019). A systematic review of domains, techniques, delivery modes and validation methods for intelligent tutoring systems. *International Journal of Advanced Computer Science and Applications*, *10*, http://dx.doi.org/10.14569/ijacsa.2019.0100312.

Stanja, J., Gritz, W., Krugel, J., Hoppe, A., & Dannemann, S. (2022). Formative assessment strategies for students' conceptions the potential of learning analytics. *British Journal of Educational Technology*, *54*, 58–75. http://dx.doi.org/10.1111/bjet.13288.

Stone, M., & Brooks, R. J. (1990). Continuum regression: Cross validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, *52*, 237–258. http://dx.doi.org/10.1111/j.2517-6161.1990.tb01786.x.

Sun, L., Hu, L., & Zhou, D. (2022). Programming attitudes predict computational thinking: Analysis of differences in gender and programming experience. *Computers & Education*, *181*, Article 104457. http://dx.doi.org/10.1016/j.compedu.2022.104457.

Sun, D., Ouyang, F., Li, Y., & Zhu, C. (2021). Comparing learners' knowledge, behaviours, and attitudes between two instructional modes of computer programming in secondary education. *International Journal of STEM Education*, *8*, http://dx.doi.org/10.1186/s40594-021-00311-1.

Swider-Cios, E., Vermeij, A., & Sitskoorn, M. M. (2023). Young children and screen-based media: The impact on cognitive and socioemotional development and the importance of parental mediation. *Cognitive Development*, *66*, Article 101319. http://dx.doi.org/10.1016/j.cogdev.2023.101319.

Swiss Conference of Cantonal Ministers of Education (2007). Intercantonal agreement on harmonisation of compulsory education (harmos agreement). http://edudoc.ch/record/24711.

Tai, R. H., Ryoo, J. H., Skeeles-Worley, A., Dabney, K. P., Almarode, J. T., & Maltese, A. V. (2022). (re-)designing a measure of students attitudes toward science: A longitudinal psychometric approach. *International Journal of STEM Education*, *9*, http://dx.doi.org/10.1186/s40594-022-00332-4.

Tónnsen, K.-C. (2021). The relevance of trial-and-error: Can trial-and-error be a sufficient learning method in technical problem-solving-contexts? *Techne serien - Forskning I Slöjdpedagogik Och Slöjdvetenskap*, *28*, 303–312, https://journals.oslomet.no/index.php/techneA/article/view/4391.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, *5*(99), http://dx.doi.org/10.2307/3001913.

UNESCO Institute for Statistics (2012). International standard classification of education: Isced 2011. *Comparative Social Research*, *30*, http://dx.doi.org/10.15220/978-92-9189-123-8-en.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual: (Python Documentation Manual Part 2)*. Scotts Valley, CA: CreateSpace Independent Publishing Platform, https://books.google.ch/books?id=KIybQQAACAAJ.

Vlachogianni, P., & Tselios, N. (2021). Perceived usability evaluation of educational technology using the system usability scale (SUS): A systematic review. *Journal of Research on Technology in Education*, *54*, 392–409. http://dx.doi.org/10.1080/15391523.2020.1867938.

Vomlel, J. (2004). Building adaptive tests using Bayesian networks. *Kybernetika*, *40*, 333–348, https://www.kybernetika.cz/content/2004/3/333.

Voronina, L. V., Sergeeva, N. N., & Utyumova, E. A. (2016). Development of algorithm skills in preschool children. *Procedia - Social and Behavioral Sciences*, *233*, 155–159. http://dx.doi.org/10.1016/j.sbspro.2016.10.176.

Vujičić, L., Jančec, L., & Mezak, J. (2021). Development of algorithmic thinking skills in early and preschool education. In *EDULEARN21 proceedings 13th international conference on education and new learning technologies* (pp. 8152–8161). IATED, http://dx.doi.org/10.21125/edulearn.2021.1650.

Vygotsky, L. S. (1978). Mind in Society: Development of Higher Psychological Processes. Harvard University Press, http://dx.doi.org/10.2307/j.ctvjf9vz4.

Wahyuningsih, S., Nurjanah, N., Rasmani, U., Hafidah, R., Pudyaningtyas, A., & Syamsuddin, M. (2020). Steam learning in early childhood education: A literature review. *International Journal of Pedagogy and Teacher Education*, *4*, 33–44. http://dx.doi.org/10.20961/ijpte.v4i1.39855.

Wang, X., Dai, M., & Mathis, R. (2022). The influences of student- and school-level factors on engineering undergraduate student success outcomes: A multi-level multi-school study. *International Journal of STEM Education*, *9*, http://dx.doi.org/10.1186/s40594-022-00338-y.

Wang, M.-T., Guo, J., & Degol, J. L. (2019). The role of sociocultural factors in student achievement motivation: A cross-cultural review. *Adolescent Research Review*, *5*, 435–450. http://dx.doi.org/10.1007/s40894-019-00124-y.

Wang, J., & Hejazi Moghadam, S. (2017). Diversity barriers in k-12 computer science education: Structural and social. In *Proceedings of the 2017 ACM SIGCSE technical symposium on computer science education*. ACM, http://dx.doi.org/10.1145/3017680.3017734.

Webb, M., Davis, N., Bell, T., Katz, Y. J., Reynolds, N., Chambers, D. P., et al. (2017). Computer science in k-12 school curricula of the 2lst century: Why, what and when? *Education and Information Technologies*, *22*, 445–468. http://dx.doi.org/10.1007/s10639-016-9493-x.

Weintrop, D., Rutstein, D. W., Bienkowski, M., & McGee, S. (2021). Assessing computational thinking: An overview of the field. *Computer Science Education*, *31*, 113–116. http://dx.doi.org/10.1080/08993408.2021.1918380.

Wickey da Silva Garcia, F., Ronaldo Bezerr. Oliveira, S., & da Costa Carvalho, E. (2022). Application of a teaching plan for algorithm subjects using active methodologies: An experimental report. *International Journal of Emerging Technologies in Learning (Ijet)*, *17*, 175–207. http://dx.doi.org/10.3991/ijet.v17i07.28733.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*, 209–212. http://dx.doi.org/10.1080/01621459.1927.10502953.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*, 33–35. http://dx.doi.org/10.1145/1118178.1118215.

Wing, J. M. (2014). Computational thinking benefits society. *40th anniversary blog of social issues in computing*, *2014*, 26, http://socialissues.cs.toronto.edu/index.html?p=279.html.

Wing, J. M. (2017). Computational thinking's influence on research and education for all. *Italian Journal of Educational Technology*, *1*, http://dx.doi.org/10.17471/2499-4324/922.

Wohl, B., Porter, B., & Clinch, S. (2015). Teaching computer science to 5-7 year-olds: An initial study with scratch, cubelets and unplugged computing. In *Proceedings of the workshop in primary and secondary computing education*. ACM, http://dx.doi.org/10.1145/2818314.2818340.

Wu, L. (2019). Student model construction of intelligent teaching system based on Bayesian network. *Personal and Ubiquitous Computing*, *24*, 419–428. http://dx.doi.org/10.1007/s00779-019-01311-3.

Xing, W., Li, C., Chen, G., Huang, X., Chao, J., Massicotte, J., et al. (2020). Automatic assessment of students' engineering design performance using a Bayesian network model. *Journal of Educational Computing Research*, *59*, 230–256. http://dx.doi.org/10.1177/0735633120960422.

Yadav, A., Mayfield, C., Zhou, N., Hambrusch, S. E., & Korb, J. T. (2014). Computational thinking in elementary and secondary teacher education. *ACM Transactions on Computing Education*, *14*, 1–16. http://dx.doi.org/10.1145/2576872.

Yates, F. (1934). Contingency tables involving small numbers and the $\chi^2$ test. *Supplement to the Journal of the Royal Statistical Society*, *1*(217), http://dx.doi.org/10.2307/2983604.

Zapata-Cáceres, M., Martín-Barroso, E., & Román-González, M. (2020). Computational Thinking Test for Beginners: Design and Content Validation. In *2020 IEEE global engineering education conference* (pp. 1905–1914). http://dx.doi.org/10.1109/EDUCON45650.2020.9125368.

Zdaniuk, B. (2014). Ordinary least-squares (OLS) model. In *Encyclopedia of quality of life and well-being research* (pp. 4515–4517). Springer Netherlands, http://dx.doi.org/10.1007/978-94-007-0753-5_2008.