Valentina Dagienė, Daranee Lehtonen, Esa Satomaa, and Mikko-Jussi Laakso

COMATH: Development and Validation of a Cross-National Assessment Instrument for Computational Thinking in Primary and Secondary Education

2026

https://doi.org/10.15388/infedu.2602.022

Publisher's PDF

Dagienė, Valentina, et al. "COMATH: Development and Validation of a Cross-National Assessment Instrument for Computational Thinking in Primary and Secondary Education." Informatics in Education, vol. 25, no. 1, 2026, pp. 59–108. https://doi.org/10.15388/infedu.2602.022.

Informatics in Education, 2026, Vol. 25, No. 1, 59–108

DOI: https://doi.org/10.15388/infedu.2602.022

COMATH: Development and Validation of a Cross-National Assessment Instrument for Computational Thinking in Primary and Secondary Education

Valentina DAGIENĖ[1] [0000-0002-3955-4751], Daranee LEHTONEN[2] [0000-0002-9695-0056], Esa SATOMAA[2] [0009-0002-2122-7928], Mikko-Jussi LAAKSO[2] [0000-0001-9163-2676]

[1]Vilnius University Institute of Educational Sciences, Vilnius, Lithuania

[2]University of Turku, Turku Research Institute for Learning Analytics, Turku, Finland

E-mail addresses: valentina.dagiene@mif.vu.lt; daranee.lehtonen@utu.fi; esa.v.satomaa@utu.fi; milaak@utu.fi

Abstract. Computational Thinking (CT) is widely recognised as a transversal competence essential for learning, problem solving, and knowledge transfer across disciplines. However, its effective integration into school education remains strongly dependent on the availability of assessment instruments that are pedagogically meaningful, psychometrically sound, and applicable across diverse educational contexts. This paper presents COMATH, a cross-national assessment instrument designed to evaluate CT in students aged 9–14. The instrument adopts a phase-based development and validation framework that integrates Bebras-inspired tasks, Item Response Theory, factor-analytic methods, learning analytics, and teacher and student feedback. The assessment was iteratively developed and piloted between 2023 and 2025 in six European countries, with data collected from 6,480 students and 155 teachers. The findings demonstrate that a phased assessment approach enables systematic calibration of task difficulty, robust evaluation of item functioning, and meaningful interpretation of student performance across age groups and national contexts. The results further highlight how well-designed CT assessment can support instructional decision-making rather than serve solely as a summative measure. The study argues for conceptualising CT assessment as a dynamic and iterative process that links measurement, psychometric validation, and pedagogical use in school education.

Keywords: Computational Thinking, Assessment Instrument, COMATH, Cross-National Assessment, Item Response Theory, Factor Analysis, Test Reliability and Validity

1. Introduction

Computational Thinking (CT) is a foundational competence in contemporary education, equipping students with the ability to approach problems systematically and to develop

robust analytical and problem-solving skills (Bilbao et al., 2023; Bocconi et al., 2022; Dagienė et al., 2021; Dagienė et al., 2024; Hsu et al., 2018). CT supports learners in structuring problems, identifying patterns, designing solution strategies, and reasoning abstractly, thereby fostering coherent and transferable learning experiences, particularly within STEM education (Dolgopolovas & Dagienė, 2024).

Despite broad agreement on its importance, CT remains a complex and multidimensional construct. Its definition and categorisation vary across the literature, reflecting the interaction of multiple cognitive processes rather than a single, well-delimited skill. Building on recent syntheses (Ezeamuzie & Leung, 2022; Shin et al., 2022), this study adopts the definition proposed by Shute et al. (2017), which conceptualises CT as a general cognitive competence applicable beyond specific programming or computing contexts (Armoni, 2016). This perspective is particularly suitable for the present study, as it supports the design and annotation of assessment tasks intended for use across school subjects and educational systems.

According to Shute et al. (2017), CT comprises six interrelated components: decomposition, abstraction, algorithms, debugging, iteration, and generalisation. Decomposition involves breaking complex problems into manageable parts, while abstraction focuses on identifying essential features through data collection, pattern recognition, and modelling. Algorithmic thinking concerns the development of structured solution steps, including considerations of efficiency, automation, and parallelism. Debugging and iteration emphasise testing, error identification, and solution refinement, whereas generalisation enables the transfer of CT skills across contexts and problem domains. Together, these components form a coherent framework for understanding and assessing CT as a higher-order cognitive competence.

As CT has gained recognition as a key 21st-century competence, the need for valid, reliable, and pedagogically meaningful assessment approaches has become increasingly evident. CT assessment is a rapidly evolving field, benefiting from advances in psychometric modelling and learning analytics, particularly through the application of Item Response Theory (IRT). The work of the CT&MathABLE project[1] demonstrates that it is possible to move beyond purely summative measurement and develop assessment instruments that support both rigorous evaluation and instructional decision-making. Well-designed CT assessments play a crucial role in ensuring that learners have equitable opportunities to develop and demonstrate their CT capabilities.

In response to this need, the CT&MathABLE research consortium initiated the development of a dedicated CT assessment instrument. Grounded in established theoretical frameworks and prior empirical research, the goal was to design a tool capable of capturing core CT competencies across different developmental stages while remaining applicable in diverse educational and cultural contexts. Particular attention was given to age appropriateness, cross-national comparability, and alignment with classroom practice.

This effort resulted in the development of the COMATH CT assessment instrument, which

[1]Computational Thinking and Mathematical Problem Solving, an Analytics Based Learning Environment, co-funded by Erasmus+ Programme (2022-1-LT01-KA220-SCH-000088736)

was piloted in six European countries: Finland, Hungary, Lithuania, Spain, Sweden, and Türkiye. The instrument targets students aged 9 to 14, a critical period during which foundational CT skills are introduced and progressively consolidated within formal education. To reflect developmental differences, three age-specific versions of the assessment were designed: COMATH1 for students aged 9–10, COMATH2 for students aged 11–12, and COMATH3 for students aged 13–14. While the underlying CT constructs remain consistent across versions, item content, linguistic complexity, and cognitive demands were systematically adapted to ensure valid and meaningful measurement at each age level.

The findings from the pilot studies provide robust empirical support for the measurement quality of the COMATH assessment and informed subsequent refinements to item design and test structure. In doing so, the study contributes to the broader objective of the CT&MathABLE project: the development of a scalable, equitable, and theoretically grounded assessment framework for CT in European school education.

This study addresses the following research questions:

RQ1: How can age-appropriate assessment tasks be systematically designed to validly measure CT skills in students aged 9–14 across diverse educational contexts?

RQ2: To what extent does the COMATH CT-assessment instrument demonstrate structural validity, reliability, and appropriate item functioning across age groups?

RQ3: How consistently does the COMATH CT assessment function across countries in terms of fairness, comparability, and applicability in diverse educational settings?

To address these questions, a retrospective and empirical analysis was conducted, drawing on the development process, documented communications, and empirical data generated within the CT&MathABLE project. By synthesising evidence from the design, implementation, and evaluation of the COMATH CT-assessment across six European countries, the study aims to inform future research and practice in the design of valid, reliable, and equitable CT assessment instruments for learners aged 9–14.

2. Review of existing Computational Thinking assessment instruments

The assessment of CT has emerged as a major priority in international education, particularly amid rapid digital transformation and the development of key 21st-century competencies (Bilbao et al., 2025). CT encompasses far more than learning to program; it involves a range of cognitive skills, including abstraction, decomposition, pattern recognition, and algorithm design. These skills enable individuals to approach problems in a logical, structured, and efficient way. Moreover, CT competencies are highly transferable across disciplines, making them valuable for enhancing learning and problem-solving in diverse educational domains.

Assessment plays a pivotal role in the successful integration of CT into K–12 education. A wide range of approaches has been used to evaluate CT skills (Kalelioglu et al., 2016; McMillan & Hellsten, 2010). Some studies rely on selected-response and constructed-

response instruments, such as the paper-and-pencil test for computer science knowledge and CT skills developed by Shell and Soh (2013), or the mixed-format assessment targeting everyday problem-solving proposed by Chen et al. (2017). In addition, researchers have examined the reliability and validity of CT assessment instruments, such as the self-efficacy scale developed by Gülbahar, Kert, and Kalelioğlu (2019) and the interview-based framework proposed by Weintrop et al. (2016).

A systematic literature review was conducted in accordance with the seven-stage model proposed by Fink (2019). The initial stages focused on clearly defining the research questions, identifying relevant bibliographic databases, and selecting appropriate search terms. A search of peer-reviewed literature published within the last decade was then conducted to identify studies that assessed CT in school education. The initial search yielded 285 articles. After removing duplicate records, 160 unique articles remained for subsequent screening and analysis. These steps were further supported by consultations with domain experts, which helped refine the scope of the review and ensure the relevance and robustness of the search strategy. In addition, the review built on and extended the systematic literature review by Erola and Mirel (2023), broadening the scope to include studies conducted at the upper secondary education level.

The selected papers were screened and included based on predefined inclusion criteria for language, publication period, and relevance. This process resulted in the exclusion of 125 articles, leaving 22 studies after title and abstract review. After full-text screening, nine articles were included in the final review, covering eleven CT assessment instruments (Table 1).

Table 1. Nine articles were included based on Erola and Mirel's (2023) systematic literature review

Authors	Publication year	Article's title
Basu, Rutstein, Xu, Wang & Shear	2023	A Principled approach to designing computational thinking concepts and practices assessments for upper elementary grades
Chen, Shen, Barth-Cohen, Jiang, Huang & Eltoukhy	2017	Assessing elementary students' computational thinking in everyday reasoning and robotics programming
Gane, Israel, Elagha, Yan, Luo & Pellegrino	2021	Design and validation of learning trajectory-based assessments for computational thinking in upper elementary grades
Kong & Wang	2021	Item response analysis of computational thinking practices: Test characteristics and students' learning abilities in visual programming contexts
Li, Xu & Liu	2021	Development and validation of computational thinking assessment of Chinese elementary school students
Relkin, de Ruiter & Bers	2020	"TechCheck": Development and validation of an unplugged assessment of computational thinking in early childhood education

Tsarava, Moeller, Román-González, Golle, Leifheit, Butz & Ninaus	2022	A cognitive definition of computational thinking in primary education.
Zapata-Cáceres, Martín-Barroso & Román-González	2021	Collaborative game-based environment and assessment tool for learning computational thinking in primary school: A Case study
Zhong, Wang, Chen & Li	2016	An Exploration of three-dimensional integrated assessment for computational thinking

The final stages of the review process focused on conducting the literature review and synthesising the findings. These stages involved evaluating the review procedure, ensuring its reliability, and systematically analysing the collected studies. In this study, analysis was carried out at the content level. The initial examination considered the target age groups or grade levels, the types and number of tasks included in each assessment, and the modes of implementation (e.g., digital or paper-based). Following this descriptive overview, attention was directed to the psychometric characteristics of the instruments, with particular emphasis on reliability and validity.

The systematic literature review and content analysis identified several key trends in CT assessment research. Most of the reviewed publications were published after 2020, indicating growing research interest in recent years. Most assessment instruments targeted elementary school students aged 5–12 and were primarily computer-based. Six of the eleven identified instruments used multiple-choice formats, while the remaining instruments employed open-ended tasks, problem-solving activities, or coding-based questions. Overall, reported validity evidence was moderate. Content validity was most commonly addressed through expert review, pilot testing, and alignment with established assessment frameworks. In contrast, construct validity was less consistently supported: although factor analyzes were conducted for eight instruments, hypothesis testing was reported in only four cases. Internal consistency was generally satisfactory, with eight instruments reporting Cronbach's alpha coefficients, most of which were high, except for one instrument with a low alpha value (0.48). Reliability evidence was the weakest overall, as only two instruments, reported within a single publication, included inter-rater reliability measures using Cohen's kappa.

Despite significant progress, several challenges remain. One key issue is the need to define clear, valid, and operational CT constructs that can guide item development and interpretation of results. Another challenge lies in integrating assessment seamlessly into everyday teaching practice, ensuring that it is perceived not as an added burden but as a meaningful tool for supporting learning. Gulbahar et al. (2025) address these concerns by proposing flexible and context-sensitive instruments aligned with instructional objectives.

Ethical considerations are also central to CT assessment. The increasing use of digital technologies raises concerns related to data privacy, algorithmic bias, and student autonomy. Consequently, assessment systems must be transparent, accountable, and respectful of learners' rights.

3. Development of COMATH

Within the CT&MathABLE project, the COMATH assessment instrument was developed as a comprehensive tool for evaluating students' CT skills in primary and lower secondary education. The overarching objective of COMATH is to provide an open-access assessment instrument with robust psychometric properties that can be applied consistently and reliably across diverse cultural and educational contexts.

Although the COMATH framework was originally designed to assess both CT and algebraic thinking, the present study focuses exclusively on the CT assessment. The analyzes reported here examine CT-related constructs, test items, and psychometric properties in detail, enabling a focused investigation of CT assessment design and validation.

The COMATH CT-assessment targets students aged 9 to 14 and is structured into three age-specific versions to reflect developmental differences in cognitive and digital skill progression:

- COMATH1 for students aged 9–10,
- COMATH2 for students aged 11–12, and
- COMATH3 for students aged 13–14.

The decision to develop separate versions of the COMATH assessment for each age group was motivated by well-established differences in students' developmental and learning trajectories. Learners in the younger age groups (COMATH1 and COMATH2) are typically at an early stage of acquiring foundational CT skills, whereas students in the older age group (COMATH3) can engage with more complex tasks that demand higher levels of abstraction, reasoning, and problem-solving. Consequently, each COMATH version was carefully designed with age-appropriate content, calibrated levels of complexity, and differentiated cognitive demands to ensure valid and meaningful measurement of students' CT abilities.

The development of the COMATH instrument was carried out at the Turku Research Institute for Learning Analytics, University of Turku, Finland, in close collaboration with Vilnius University, Lithuania, and other project partners. To ensure strong validity and reliability, the development process was grounded in established theoretical and design frameworks and informed by interdisciplinary expertise. Researchers from multiple academic disciplines and cultural contexts worked alongside practising teachers, whose classroom-based perspectives contributed to the instrument's pedagogical relevance and practical applicability.

The design process integrated both quantitative and qualitative research methods, including iterative item development, expert evaluation, and systematic alignment with curricular and pedagogical objectives. Following its initial development, the COMATH CT-assessment was piloted from October 2023 to the beginning of 2025 across six countries: Finland, Hungary, Lithuania, Spain, Sweden, and Türkiye. The findings from this first pilot phase serve as the empirical foundation for the present study and are used to evaluate the CT assessment component of COMATH.

Insights from the first pilot study informed targeted refinements to the instrument, which were implemented before a larger-scale pilot. This subsequent phase focuses on further examining and confirming the validity and reliability of the finalised COMATH CT-assessment across diverse educational contexts.

For the development of the test items, we utilised existing CT assessment instruments identified in the systematic review and demonstrated at least moderate psychometric quality. The CT test items were supplemented with Bebras tasks (mainly selected from the 2022 challenge), which are widely recognised as reliable tools for CT skills. We also generated new test items when no existing items matched the CT skills intended for assessment. Given the different developmental phases of the target students, the assessment instrument's difficulty level should reflect the varying complexities of students' digital skill development and capture the levels of their CT skills. Therefore, the COMATH was tailored to three distinct age groups.

3.1 Item Response Theory Analysis of Bebras Tasks

To support the development of the COMATH assessment, we conducted an Item Response Theory (IRT) analysis of tasks from the Bebras Challenge 2022 (www.bebras.org). Bebras tasks are designed to engage students with Informatics and CT concepts and to stimulate problem-solving and algorithmic reasoning (Araujo et al., 2019; Dagienė & Sentence, 2016). Our objective was to identify tasks suitable for integration into the COMATH assessment, particularly those that effectively differentiate students according to their CT ability levels.

The analysis was based on responses from 88,041 students in Lithuania and Hungary. Item–Person Maps and Item Characteristic Curves (ICCs) were used to evaluate how well tasks aligned with students' ability distributions and how effectively they differentiated between lower- and higher-performing learners. Item difficulty and discrimination parameters were analyzed across age groups to evaluate task functioning at the item level. In addition, percentile norms were calculated separately for each country to support the interpretation of individual student performance.

As an example, the suitability of Bebras tasks for students aged 8–10 was examined using an Item–Person Map (Figure 1). This visualization presents estimated student ability levels (θ) on the horizontal axis and the number of students at each level on the vertical axis. For this age group, ability estimates ranged approximately from −2 to slightly above 2, providing a reference for assessing how well task difficulty aligned with students' skill levels.

Figure 1. Item-Person Map of a Bebras task for Lithuanian and Hungarian students aged 8–10

The discriminative performance of individual tasks across the ability continuum was also examined (Figure 2). The results showed that some tasks were not optimally targeted for this age group. For example, task 2022-SK-03 was too difficult, providing limited differentiation among students with lower ability levels. In contrast, task 2022-IN-01 was too easy, providing limited differentiation among students with higher ability levels.

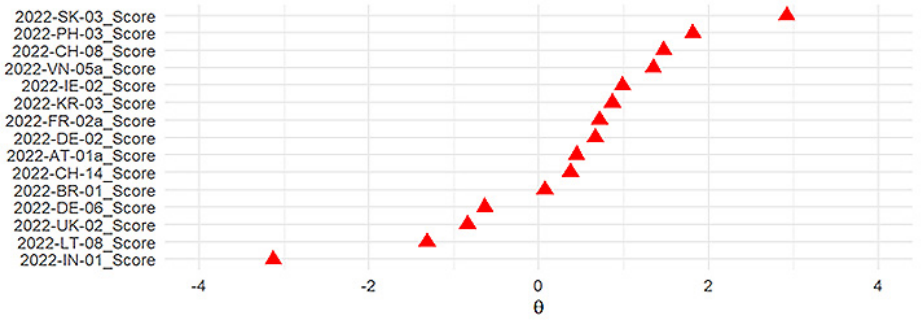


Figure 2. Difficulty parameter estimates from the IRT analysis of the Bebras tasks for Lithuanian and Hungarian students aged 8–10

To further examine the discriminatory properties of the tasks, ICCs were employed to model the relationship between students' ability levels (θ) and the probability of a correct response, P(θ). Each task was represented by its own ICC, illustrating how response probability varied across the ability continuum (Figure 3).

Within the IRT framework, the discrimination parameter (a) indicates how effectively a task differentiates between students at different ability levels. Higher discrimination values, reflected in steeper ICCs, signal greater sensitivity to differences in ability. In this analysis, tasks such as 2022-CH-08 and 2022-BR-01 showed particularly strong discriminative power.

The horizontal position of an ICC represents task difficulty: curves located further to the right correspond to more challenging tasks, whereas those shifted to the left indicate easier ones. The lower asymptote reflects the probability of guessing.

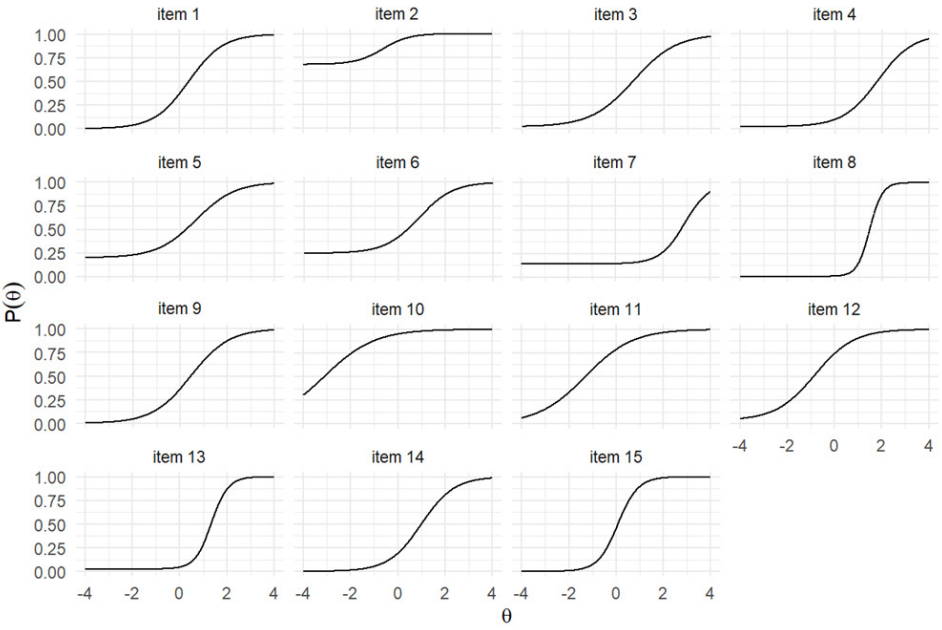


Figure 3. ICCs for Bebras tasks administered to Lithuanian and Hungarian students aged 8–10.

Item order: 1 = 2022-CH-14, 2 = 2022-DE-06, 3 = 2022-FR-02a, 4 = 2022-PH-03, 5 = 2022-DE-02, 6 = 2022-KR-03, 7 = 2022-SK-03, 8 = 2022-CH-08, 9 = 2022-AT-01a, 10 = 2022-IN-01, 11 = 2022-LT-08, 12 = 2022-UK-02, 13 = 2022-VN-05a, 14 = 2022-IE-02, 15 = 2022-BR-01).

Based on these results, 41 Bebras tasks with strong discrimination were selected for potential inclusion in the COMATH CT assessment. Their difficulty parameters (b) were then examined to ensure balanced coverage across the ability range. Tasks with $b > 1$ were considered difficult, those with $b < -1$ were considered easy, and those between approximately -0.5 and 0.5 were classified as moderate, supporting a balanced assessment of the target population.

3.2 Expert Evaluation

In addition to the IRT analyzes, the content validity of the selected Bebras tasks was examined to confirm their suitability and representativeness for assessing CT skills in the development of the COMATH test items. Content validity concerns the degree to which an assessment instrument adequately captures the intended construct for its defined purpose (Almanasreh et al., 2019).

To evaluate content validity, 13 CT experts from the project partner countries were consulted. These experts had, on average, 17 years of experience in CT education. They were

asked to judge the extent to which each selected task reflected specific CT components, based on the classification proposed by Shute et al. (2017). Ratings were provided using a three-point scale ("Well", "Somewhat", or "Not at all"). Based on these evaluations, the content validity ratio (CVR) was calculated using the following formula:

$$\mathrm{CVR} = \frac{(n_e - N/2)}{N/2}$$

in which n_e represents the number of experts who rated the task as "Well = 1" or "Somewhat = 0.5" and N is the total number of experts (13). The CVR, originally introduced by Lawshe (1975), is a widely used quantitative indicator of content validity. CVR values range from −1 to 1, with higher values reflecting stronger agreement among experts regarding the relevance and necessity of an item. Statistical significance is determined using Lawshe's critical values; for a panel of 13 experts, a CVR exceeding 0.54 indicates acceptable content validity at the 0.05 significance level.

All thirteen CT experts participated in evaluating the selected tasks and provided feedback on several dimensions, including item clarity and comprehensibility, alignment with the intended CT skills, appropriateness for the target age groups, and suggestions for improvement.

The CVR results indicated that the majority of tasks primarily assessed algorithmic thinking, with some tasks also addressing abstraction. This finding is consistent with earlier research by Araujo et al. (2019), which showed that Bebras Challenge tasks often combine multiple CT components, with algorithmic thinking playing a central role. Based on these results, the 41 Bebras tasks selected for COMATH were classified into two categories: tasks assessing algorithmic thinking only, and tasks requiring algorithmic thinking in combination with other CT skills.

In addition to quantitative ratings, experts offered qualitative comments concerning task content, structure, age appropriateness for students aged 9–10, 11–12, and 13–14, and potential cultural considerations. This feedback was used to further refine the assessment items. Revisions based on expert input focused on improving clarity, relevance, and completeness, thereby enhancing the suitability of the tasks for use across diverse educational contexts.

3.3 Usability Testing

The preliminary version of the assessment instrument was delivered through ViLLE, a digital learning environment developed at the University of Turku, Finland (Laakso et al., 2018). Alongside the expert review, the instrument was trialled with a small group of Finnish students to evaluate its face validity, with a particular focus on usability and clarity. A think-aloud approach was adopted, allowing students to verbalise their thoughts and actions as they completed the assessment tasks on the ViLLE platform.

Usability testing was conducted in September 2023 using a subset of the COMATH CT

items. Three students participated: a girl in Grade 4, a boy in Grade 5, and a boy in Grade 8. Each student completed the age-appropriate assessment during a session scheduled for one standard school lesson (45 minutes), with the option to finish earlier if desired. Participants were encouraged to ask questions or indicate any difficulties encountered during the test.

Completion times varied by age: the Grade 4 student finished in 28 minutes, the Grade 8 student in 34 minutes, and the Grade 5 student in 40 minutes. Overall, the assessment functioned smoothly, and no major usability problems were identified. However, the amount of verbal feedback provided was limited, possibly because the students were unfamiliar with one another and with the facilitator, which may have reduced their willingness to ask questions or comment on specific items.

Despite the limited volume of feedback, several useful observations emerged. The youngest participant initially felt uncertain about the procedure for submitting responses, suggesting the need for clearer instructions at this stage of the assessment. In addition, some tasks with relatively dense textual content were perceived as demanding for younger students, despite efforts to minimise reading load. Nevertheless, the overall test structure and duration were found to be appropriate, and the assessment was considered ready for implementation in the subsequent project pilot study.

3.4 Finalising the COMATH

In the final phase of the first development, the selected items were organised into a coherent assessment framework designed to measure students' CT skills. Each test was structured to be completed within 40–45 minutes, corresponding to the typical duration of a school lesson across partner countries. This time allocation was selected to ensure practical feasibility within regular timetables while preserving sufficient depth and measurement quality.

After incorporating expert feedback, usability critiques, and IRT analysis, we modified existing Bebras tasks to make them less complex. This included shortening text passages and reducing the number of multiple-choice answers. We also adjusted tasks for different age groups; for instance, we repurposed a medium-difficulty task meant for older participants as a difficult task for younger participants, and vice versa. Additionally, we designed entirely new tasks inspired by test items identified in our systematic literature review. These efforts ensured that the tasks developed align with the test's objectives while catering to the diverse skill levels of various age groups.

The developed test was implemented on the ViLLE digital platform and translated into the eight official languages of the participating countries: Basque, Catalan, Finnish, Hungarian, Lithuanian, Spanish, Swedish, and Turkish. These translations ensured accessibility for students from various educational backgrounds and facilitated consistent administration during the first pilot phase in each country.

The selected tasks were subsequently organised into two categories: (1) tasks assessing algorithmic thinking exclusively and (2) tasks combining algorithmic thinking with addi-

tional CT components. Within each category, items were positioned along a common difficulty continuum based on IRT estimates for each age group. The difficulty scale ranged from −2 to 2 for each COMATH level, ensuring that tasks were appropriately aligned with the expected ability range of the target students.

After grouping tasks by difficulty level, we selected one task with the strongest discrimination properties from each group to include in the COMATH assessment. However, as illustrated in Figures 4–6, the resulting set of tasks did not provide consistent coverage across the entire range of ability levels within each age group. A well-designed assessment should include items that span a broad spectrum of difficulty, so additional tasks were developed to address these gaps and ensure a more balanced distribution of easy, moderate, and challenging items across the difficulty continuum.

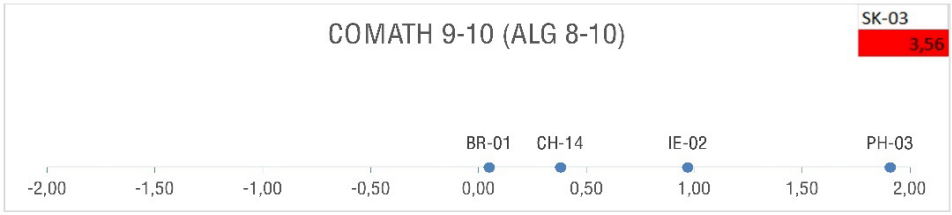


Figure 4. Difficulty of the algorithmic thinking tasks for students aged 9–10

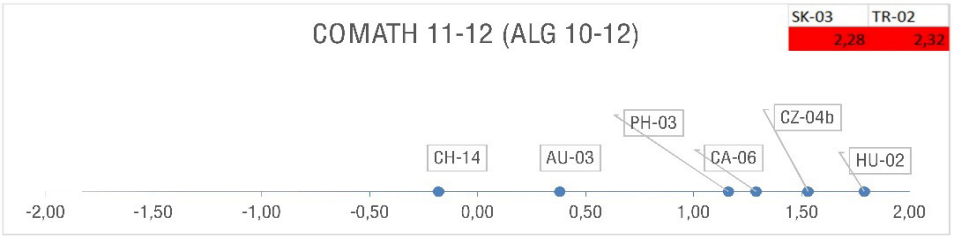


Figure 5. Difficulty of the algorithmic thinking tasks for students aged 11–12

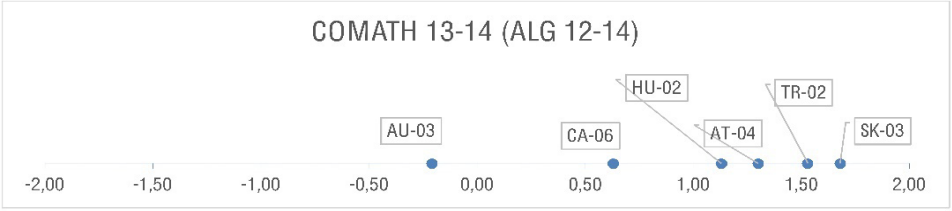


Figure 6. Difficulty of the algorithmic thinking tasks for students aged 13–14

The development of supplementary tasks was informed by analyzes of tasks from the Bebras Challenge and by findings from the systematic literature review. New tasks were created through several strategies: adapting existing Bebras tasks by simplifying textual

descriptions or reducing the number of response options to lower task difficulty; reallocating tasks across age groups by assigning medium-difficulty tasks from older groups as more challenging tasks for younger students, and vice versa; and designing entirely new tasks inspired by CT assessment instruments identified in prior research.

For the first pilot study, a total of 29 CT test items were included across COMATH 1–3. Of these, 14 tasks focused exclusively on algorithmic thinking, while 15 assessed it alongside other CT components. Most tasks (23 items) were developed in two parallel versions (A and B) that were conceptually equivalent but differed in surface features, such as images, visual orientation, and minor textual variations. These parallel forms were intended to reduce test-form effects and to support subsequent analyzes of item equivalence and refinement. In addition, 18 tasks served as anchor items, with eight tasks included across all three age levels to support comparability.

Each assessment item was administered in one of two parallel versions, and students were randomly assigned to complete either version A or B (see Figure 7), ensuring balanced exposure across countries and age groups. Each age-specific test consisted of 18–19 items, providing sufficient coverage of the CT construct while keeping the testing duration appropriate for school settings. The task set was organized into two categories: items assessing algorithmic thinking alone and items integrating algorithmic thinking with other CT skills, such as abstraction, decomposition, and pattern recognition. This structure allowed algorithmic thinking to be examined both as a central CT component and in relation to broader computational processes.

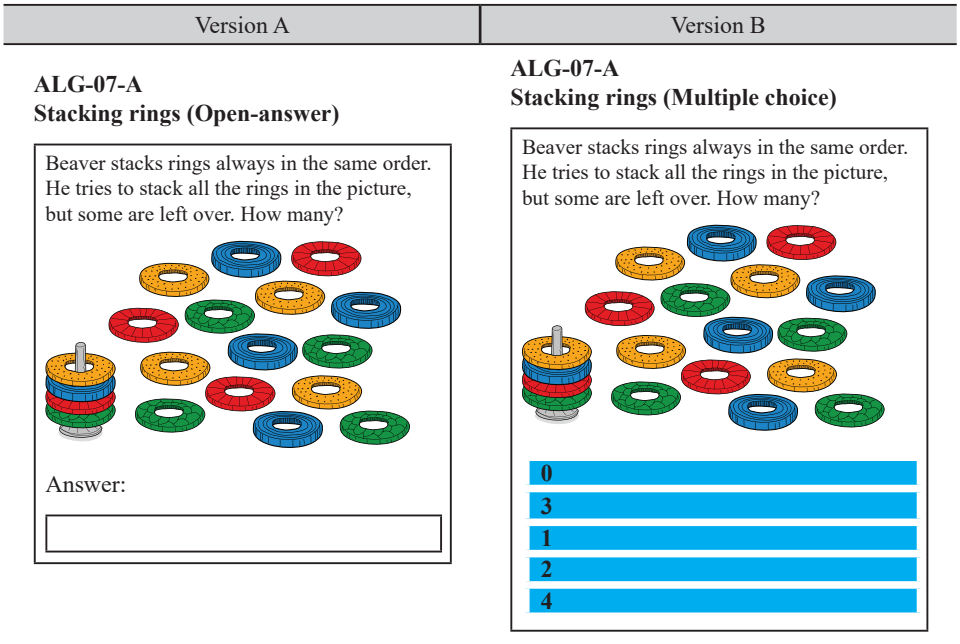


Figure 7. Example of a task A with an alternative version B.

Following data collection from the first pilot, extensive data cleaning and pre-processing were conducted, resulting in a final dataset comprising responses from 3,350 students and 53,268 valid item-level observations. All analyzes were performed separately for each age group to account for developmental differences and to support age-specific validation. The psychometric evaluation of the COMATH assessment employed a multi-method approach, including Exploratory Factor Analysis, Confirmatory Factor Analysis, Item Response Theory modelling, and reliability analyzes. Together, these methods provided complementary evidence regarding the instrument's structural validity, reliability, and item-level performance.

The assessment items were systematically designed to be developmentally appropriate and cognitively demanding at the intended level, incorporating CT components aligned with the learning objectives specified in the national curricula of the participating countries. Before the pilots, the project team thoroughly reviewed all items to eliminate typographical errors and resolve technical issues, ensuring the digital assessment environment was accurate, clear, and functionally reliable.

4. Methodology

Despite extensive research, assessing CT remains a substantial methodological challenge (Tang et al., 2020). Unlike traditional academic domains, CT encompasses a set of interconnected cognitive processes that do not readily align with conventional standardised testing formats. As a result, researchers have proposed a range of theoretical frameworks and specialised assessment tools to capture the multidimensional and latent nature of CT competences (Bilbao et al., 2025).

The development of the COMATH task set, described in the preceding section, was guided by these methodological considerations. The process was theoretically grounded, interdisciplinary, and informed by both empirical research and classroom practice. As a result, the task set was designed to support rigorous psychometric analysis while remaining accessible, engaging, and meaningful for students across different age groups and cultural contexts.

To empirically evaluate the quality of the CT component of the COMATH assessment, two pilot studies were conducted. The first pilot focused on examining initial item functioning, underlying construct structure, and the feasibility of administering the assessment across participating countries. Findings from this phase informed targeted refinements to selected items and test structures. The second pilot extended this evaluation by further examining the instrument's psychometric properties, including factor structure, item parameters, and reliability across age groups. Together, these pilot studies provide the empirical foundation for the methodological analyzes presented in this paper.

Beyond psychometric considerations, contemporary CT assessment research highlights the importance of pedagogical relevance and contextual validity. Several studies emphasise the active involvement of teachers in both the development and interpretation of as-

sessment instruments. Teacher participation in task co-design and analysis of assessment results strengthens alignment with curricular objectives, facilitates meaningful classroom integration, and supports the identification of recurring learning difficulties that can inform targeted instructional interventions.

4.1 Item Response Theory Analysis

Item Response Theory (IRT) represents one of the most advanced and widely adopted psychometric approaches for assessing CT. IRT models latent abilities by examining the probabilistic relationship between students' item responses and their underlying competence levels. Unlike classical test theory, which focuses primarily on total test scores, IRT emphasises item-level characteristics, such as difficulty and discrimination, thereby enabling a more fine-grained analysis of both task functioning and learner performance. This item-centred perspective makes IRT particularly well-suited for evaluating complex, multidimensional constructs such as CT, where performance is influenced by varying cognitive demands and problem-solving strategies.

The application of IRT in CT assessment offers several important advantages. First, it supports the development of precise and scalable assessment instruments that are sensitive to differences in students' competence levels. Second, IRT facilitates the construction of calibrated item banks and longitudinal assessment frameworks, enabling the monitoring of learning progression over time. Third, because item parameters are estimated independently of the tested sample, IRT enhances the fairness, comparability, and interpretability of assessment results across diverse learner populations and educational contexts.

From a psychometric standpoint, IRT has been increasingly adopted to strengthen the validity and reliability of CT assessment instruments. Gyamfi and Acquaye (2023) emphasise that IRT allows for stable estimation of student ability even when different test forms or difficulty levels are administered. Furthermore, IRT provides the methodological foundation for adaptive testing and flexible assessment designs, which are particularly valuable in heterogeneous classroom environments.

Empirical studies further demonstrate the suitability of IRT for CT assessment. For instance, Kong and Wang (2021) applied IRT models to analyze student performance in digital learning environments involving algorithmic reasoning and complex problem-solving tasks. Their results show that multi-step items with higher cognitive demands tend to exhibit stronger discrimination, indicating their effectiveness in distinguishing between different levels of CT competence. Such findings highlight the importance of item-level analysis for identifying the tasks most appropriate for assessing specific CT components.

In the present study, IRT analysis was employed to evaluate the difficulty and discrimination properties of the COMATH test items. The two parallel item versions (A and B) were analyzed separately. A two-parameter IRT model was applied, estimating one parameter for item difficulty and one for item discrimination for each test item.

4.1.1 Difficulty Parameter

The difficulty parameter indicates how challenging a test item is. Items with values close to 0 are of moderate difficulty, while negative values represent easier items (lower values indicate lower difficulty) and positive values represent harder items (higher values indicate greater difficulty). For example, an item with a difficulty parameter of -2 is considered very easy, whereas an item with a parameter of 2 is considered very difficult. Ideally, difficulty estimates should be spread across the range (-2, 2) to ensure a balanced mix of item difficulties. Items with values significantly beyond this range may be too easy or too difficult, potentially affecting the test's effectiveness.

In this report, items with difficulty estimates below -4 or above 4 are classified as abnormal, while those within (-4, -2) and (2, 4) are flagged as potentially too easy or too difficult, respectively. Baker (2001) and Gyamfi and Acquaye (2023) suggest that typical difficulty values range from -3 to 3, so the threshold values of -4 and 4 were chosen as a guideline, extending slightly beyond this typical range.

4.1.2 Discrimination Parameter

The discrimination parameter reflects how well a test item differentiates between students based on their ability. A value close to 0 suggests little relationship between ability level and the likelihood of answering correctly. In contrast, a high discrimination value indicates that students with greater ability are much more likely to answer correctly than those with lower ability.

Discrimination values below 0.65 suggest poor differentiation, but this is usually assessed using item characteristic curves (ICCs) rather than numerical values alone. The 0.65 threshold is based on Baker's (2001) classification, where values between 0.35 and 0.64 indicate low discrimination, and values between 0.01 and 0.34 indicate very low discrimination. A negative discrimination parameter is particularly concerning, as it suggests that higher-ability students are more likely to answer incorrectly than lower-ability students.

While high discrimination values are generally desirable, very high values may indicate instability in the model. In this report, items with discrimination estimates above 4 are classified as abnormal. Baker (2001) and Gyamfi and Acquaye (2023) suggest that typical discrimination values range from -3 to 3, so a threshold of 4 was chosen as a guideline. However, an abnormally high discrimination value does not necessarily indicate a flawed test item.

4.1.3 Summary of Problematic IRT Parameter Estimates

The following is a summary of the interpretations applied to problematic IRT parameter estimates in the subsequent analyzes:

- Items with difficulty parameter estimates in the ranges (-4, -2) or (2, 4) are flagged as potentially too easy or too difficult, respectively.

- Items with difficulty estimates below -4 or above 4 are reported separately, as such extreme values may indicate not only excessive ease or difficulty but also potential model instability.
- Items with discrimination parameter estimates between 0 and 0.65 are reported due to their poor ability to differentiate students based on ability.
- Items with discrimination estimates below 0 are highlighted, as they suggest that higher-ability students are less likely to answer correctly than lower-ability students, which contradicts the expected pattern.
- Items with discrimination estimates above 4 are reported as potentially unstable, though a high discrimination value does not necessarily indicate a flawed test item.

4.2 Test Reliability Estimation

Test reliability was assessed in the first pilot study using both Cronbach's alpha and the omega total coefficient across three factors. While Cronbach's alpha is the most commonly used measure of test reliability, some researchers advocate the omega coefficient as a more robust alternative because it makes fewer assumptions about the test items. Therefore, both measures were reported.

Additionally, the test items were evaluated using item-total correlations and item-rest correlations. These correlations measure the relationship between each item's score and the overall test score. In item-total correlation, the test item is included in the total score, whereas in item-rest correlation, it is excluded.

Factor analysis was conducted in the first pilot study to explore the underlying factor structure of the test items. Initially, an exploratory factor analysis (EFA) was used to identify the latent factor structure without imposing any prior assumptions. This process involved determining the appropriate number of factors and assessing the test items' correlations with those factors.

Subsequently, confirmatory factor analysis (CFA) was employed to investigate how well the data supported various theoretical factor structures. The factor models analyzed included a 1-factor model, where all items reflected a single underlying CT construct; a 2-factor model, in which items were divided into algorithmic thinking tasks and those requiring algorithmic thinking alongside other CT skills; and a factor model derived from the EFA.

4.3 First pilot and feedback gathering

The first pilot study was designed to evaluate how effectively the test items measured students' CT skills across different age groups and educational contexts. To this end, analyzes were conducted separately for each age group within each participating country, across all age groups within individual countries, and across the full international sample. This multilevel analytical approach enabled an examination of item functioning at both national and cross-national levels, supporting the identification of items that effectively differentiate between varying levels of student competence. In addition, the analysis provided detailed

insights into the discriminatory power of individual tasks, highlighting those items most suitable for assessing specific skill levels.

4.3.1 Data collection

The collected data comprised answers and response times from the first pilot CT test, taken by students from six different countries (Finland, Hungary, Lithuania, Spain, Sweden, and Türkiye). There were three different test versions, each designed for a specific age group: COMATH1 for ages 9–10, COMATH2 for ages 11–12, and COMATH3 for ages 13–14. Each test item also had an A and a B version, with each student completing one. The version of each test item was assigned randomly.

Each test item fell into one of two subgroups, assessing either algorithmic thinking skills (ALG) or algorithmic thinking along with other CT skills (OTH). To exclude guesswork from the analysis, 1,366 responses with response times of less than 5 seconds were removed, as the tasks were too complex to be completed in such a short time. The distribution of response times was also taken into account when determining this threshold. Additionally, 166 responses with response times exceeding 10 minutes were removed as outliers.

After these adjustments, the final dataset included 3,350 students and 53,268 valid (non-NA) responses. The number of students per country in the final dataset is shown in Table 2.

Table 2. Number of students included in the analysis of the CT test.

	Finland	**Sweden**	**Lithuania**	**Hungary**	**Turkey**	**Spain**	**All**
Age Group 1 (9–10)	95	18	787	100	78	12	1090
Age Group 2 (11–12)	107	120	635	156	110	211	1339
Age Group 3 (13–14)	63	37	552	216	28	25	921
All	**265**	**175**	**1974**	**472**	**216**	**248**	**3350**

4.3.2 Data Analysis Approach

Each age group was analyzed separately. Correlation plots of the test items were examined to identify relationships between them and to determine if any items had very weak correlations with others. Reliability assessment, factor analysis, and IRT analysis were then carried out as presented earlier.

Since each test item had an A and a B version with only minor differences, such as graphical or textual details, these versions were merged (i.e., treated as the same test item) for reliability and factor analyzes. This was necessary because no student completed both versions of the same item, so direct correlation coefficients could not be calculated, and both methods require a complete correlation matrix. Given their near-identical nature, it was reasonable to assume that students would respond similarly to them.

4.4 Further development of COMATH based on the first pilot results

Following the first pilot study, the COMATH CT-assessment instrument underwent systematic refinement to address issues identified through empirical analysis and user feedback. The further development phase was guided by multiple data sources from the first pilot, including students' performance data and feedback from students and teachers collected through additional surveys and interviews (see Lehtonen et al., 2025). This multisource evidence base enabled both psychometric and practical considerations to be incorporated into the revision process.

As a result of these revisions, the CT assessment was streamlined into a single test version per age group, replacing the parallel A and B versions used in the first pilot. The number of CT items per age group was reduced from 18–19 in the first pilot to 14 in the revised instrument. These refinements improved the practicality and psychometric robustness of the COMATH assessment and formed the basis for its evaluation in the second pilot conducted in autumn 2024 – spring 2025.

4.4.1 Completion time

The insights gained from the first pilot informed targeted modifications to the task set. Students' response times were analyzed to determine an appropriate overall test length, ensuring that the assessment could be completed within a single 40-minute lesson. IRT analyzes were used to identify tasks with suitable difficulty and discrimination parameters, aiming for a balanced distribution of items from easy to challenging across age groups. Based on these results, underperforming or redundant items were removed, and the content of selected tasks was revised to improve measurement quality and reduce unnecessary cognitive or time demands.

4.4.2 Quality analysis of test items

A comprehensive statistical analysis was conducted to evaluate the quality of the COMATH test items used in the first pilot study, to inform the refinement of the assessment for subsequent implementation. The analysis combined confirmatory factor analysis (CFA) and Item Response Theory (IRT) to examine both the structural alignment of items with the intended CT construct and their psychometric performance at the item level.

The evaluation began with CFA to examine the relationships among test items within each age group and to assess whether they consistently measured the same underlying CT construct. Items that exhibited weak correlations with other items were considered misaligned and were removed from the task set. This step ensured greater coherence of the measurement model prior to item-level analysis.

Subsequently, a two-parameter IRT analysis was applied to evaluate item discrimination and difficulty. Items with low (below 0.65) or negative discrimination were excluded because they lacked sufficient sensitivity or exhibited atypical response behaviour that could

compromise measurement validity. In a small number of cases, exceptionally high discrimination values (above 4) were observed; these items were reviewed individually, and retention decisions were based on a combination of statistical evidence and expert judgment.

Item difficulty estimates were used to position tasks along the ability scale and to evaluate coverage across the target competence range. In line with established psychometric guidelines, items with extreme difficulty values (above 3) were excluded, as they contributed little to effective measurement. The remaining items were selected to achieve a balanced distribution of difficulty levels, ensuring that the assessment could meaningfully distinguish between lower-, middle-, and higher-performing students within each age group. Where parallel versions of an item were available, the version that best supported an even difficulty distribution was retained. When difficulty distributions were skewed, selected items were revised to improve balance and coverage.

To improve overall balance in item number, content coverage, and difficulty level, items for the second pilot were carefully selected based on their discrimination values, difficulty levels, and relevance to the target content. Items with weak correlations with other test items, very low discrimination values, or difficulty levels that were too high or too low were excluded from the pool. The remaining items were calibrated and positioned along the ability scale according to their assessed difficulty values. Additional revisions were made to some of these items to refine difficulty levels, ensuring a balanced progression from easier to more challenging tasks across the test. This systematic approach aimed to optimize the test's reliability and validity while catering to the intended skill range of participants.

Figure 8 presents a development example of algorithmic thinking items tailored specifically for Age Group 1, illustrating how item difficulty and content alignment were achieved through iterative refinement.

Figure 8. Item difficulty distribution for algorithmic thinking items (Age Group 1)

NOTE: Items included in the test are shown in black (above the scale). Excluded items are shown in red (below the scale). Revised items are shown in grey boxes (above the scale). Difficulty values for revised items are based on the first pilot.

As a result of this quality analysis, the CT assessment was streamlined from 18–19 items per age group to 14 (seven items per CT category), improving feasibility for classroom implementation while preserving construct coverage. Item selection and revision decisions were guided by a combination of CFA results, IRT parameters, and content relevance, ensuring both psychometric robustness and pedagogical appropriateness. The refined task set

served as the basis for the second pilot study, which aimed to confirm the stability of the revised instrument's psychometric properties and further evaluate its performance across age groups and national contexts.

In its finalised form, the COMATH assessment comprises 24 distinct CT test items across the three age groups, with selected items serving as anchor tasks to support comparability across levels. This streamlined structure preserves psychometric robustness while ensuring that the assessment can be completed within a standard lesson duration, thereby facilitating its integration into regular school practice and enabling reliable cross-age and cross-national analyzes in the second pilot.

4.5 Second pilot: data collection and analyzes

Following the quality analysis and item refinement based on the first pilot study, the CT assessment was finalised for the second pilot. The revised CT test consists of 14 items for each age group. These items are evenly distributed across the two test sections, with seven items per section, ensuring balanced coverage of core CT skills while improving the practicality of the assessment for classroom use.

The second pilot study was conducted during autumn 2024–spring 2025 to extend and strengthen the psychometric evaluation of the COMATH assessment. Building on the findings of the first pilot and a corrected set of tasks, this phase focused on further examining the instrument's factor structure, item-level parameters, and reliability across age groups. Particular attention was given to confirming the stability of the measurement model and the consistency of item functioning following revisions introduced after the initial pilot. In addition, the second pilot enabled the assessment of the instrument's performance across different national contexts using a refined set of tasks. Together, these analyzes provide more robust evidence for the validity and reliability of the COMATH CT assessment. Statistical analyzes were done using R 4.4.2 software (R Core Team, 2024).

Primary data consisted of answers, answer correctness (did the subject answer the test item correctly), and answer times to the Pilot 2 CT test from test subjects in 6 different countries. There were 3 different test versions for 3 different age groups. Each test item belonged to one of two subgroups: algorithmic thinking (ALG) or algorithmic thinking and others (OTH). Each test had 14 test items. In Lithuanian data, some test items had multiple answers from users with the same user IDs. Here, for each user, only the earliest timestamp was retained, and the others were dropped from the data. 46 answers from Age Group 1, 1 answer from Age Group 2, and 3 answers from Age Group 3 were dropped this way. This left 39,933 non-NA responses from 3,130 subjects in the preliminary data.

To remove guessers from analysis, 571 answers with answer times shorter than 3 seconds were removed, as the tasks were too complex to do in such a short time. After this, all subjects with 70% or more missing values in their responses were excluded from the data. 85 subjects were dropped this way. Finally, all missing values were imputed with 0, assuming the primary reason for not answering the test item was not knowing how to answer. This

left 42,630 non-NA answers from 3045 subjects for the final data to be analyzed. The number of subjects by country and age group in the final dataset is shown in Table 3.

Table 3. Number of students included in the analysis of Pilot 2 CT after removing answers with too small answer times and dropping students with 70% or more missing answers.

	Finland	**Sweden**	**Lithuania**	**Hungary**	**Turkey**	**Spain**	**All**
Age Group 1 (9–10)	10	3	608	81	79	85	866
Age Group 2 (11–12)	63	50	715	58	69	103	1058
Age Group 3 (13–14)	127	136	556	114	60	128	1121
All	**200**	**189**	**1879**	**253**	**208**	**316**	**3045**

Each age group conducted a separate test, which was analyzed separately. Test reliability was tested by calculating Cronbach's alpha coefficient for each age group. However, as the variables in this case were binary variables (1 for a correct answer and 0 for a wrong answer) and regular Cronbach's alpha assumes continuous variables, a deflation-corrected alpha coefficient was also calculated for each group. This deflation-corrected alpha coefficient was calculated using Somers' Delta, which is better suited for binary variables.

Item-level performance was evaluated using both item–total correlations (RIT) and item–rest correlations (RIR). The item–total correlation represents the association between an individual item score and the overall test score, whereas the item–rest correlation measures the association between an item score and the total test score with that item excluded. In addition, deflation-corrected versions of these indices (DIT and DIR) were computed using Somers' Delta, as these measures are more appropriate for dichotomously scored items. Items with a DIR value below 0.30 were considered insufficiently consistent with the rest of the test and were removed from subsequent analyzes.

The structural validity of the CT assessment was examined through CFA. Two alternative measurement models were tested: a one-factor model in which all items were assumed to load on a single latent CT construct, and a two-factor model in which items were grouped into algorithmic thinking items and items integrating algorithmic thinking with other CT components. For each age group, two preliminary tests were also conducted to assess the suitability of the data for factor analysis: the KMO test and the Bartlett test. Data are suitable for factor analysis if the KMO test's Overall MSA is over 0.8 and the Bartlett test p-value is under 0.05. The criteria for good structural validity were good model indicators (Chi-square test, $p > .05$; Comparative Fit Index, CFI > .95; Tucker-Lewis Index, TLI > .95; Root Mean Square Error of Approximation, RMSEA < .06). It is good to note that some researchers have interpreted CFI values over 0.9 as acceptable (e.g. Bentler & Bonett, 1980). Because the variables in the analysis were binary (1 if the answer to the test item was correct, 0 if incorrect), the DWLS estimator was used with the CFA function in the lavaan package, setting ordered=TRUE in R (Rosseel, 2012). When the CFA function of the lavaan package is used with the DWLS estimator, it returns both a standard and a scaled version of the model indicators; here, the scaled indicators are reported.

IRT was used to evaluate item difficulty and discrimination. The two-parameter model estimated a difficulty (b) and a discrimination (a) parameter for each item. Interpretation of the parameter estimates followed the same approach used in Pilot 1.

In the applied two-parameter model, the guessing parameter was not estimated. Instead, for multiple-choice items, it was fixed at 1 divided by the number of response options, and for open-ended items, it was set to zero.

To support the interpretation of total test scores, score quantiles were calculated for the full sample and separately by country. For example, the 50th percentile represents median performance, while the 90th percentile identifies the top-performing 10% of students.

5. Results and discussion

This section presents the results of the psychometric analyzes conducted to evaluate the CT assessment developed within the COMATH framework. The findings are based on data from the second pilot study and focus on examining the structural properties, item-level functioning, and overall measurement quality of the instrument across age groups.

The results are organized around three complementary analytical approaches. First, test reliability is estimated using Cronbach's alpha and deflation-corrected alpha. Item-level performance is also examined using item-rest correlations, which reflect the consistency of items with the rest of the test. Second, CFA is used to assess the structural validity of the assessment and determine whether the items reflect a single latent CT construct or multiple dimensions. Third, IRT analyzes are used to estimate item difficulty and discrimination parameters, thereby evaluating how effectively individual tasks differentiate students across ability levels.

Together, these findings provide evidence regarding the validity, reliability, and practical suitability of the COMATH assessment. The results also highlight age-related differences in item functioning and test targeting, which are discussed in relation to refining the task set and the instrument's overall development.

5.1 COMATH Reliability

Internal consistency reliability was evaluated using Cronbach's alpha for each age group. Because the test items were dichotomously scored (1 = correct, 0 = incorrect), and traditional Cronbach's alpha assumes continuous variables, a deflation-corrected alpha coefficient was also calculated using Somers' Delta, which is more appropriate for binary data. Table 4 shows alphas before excluding items with DIR < 0.3.

Table 4. Cronbach's alpha and deflation-corrected alpha using Somers' Delta for Pilot 2 CT test by age group.

	Cronbach's alpha	Deflation- corrected alpha
Age Group 1 (9–10)	0.71	0.85
Age Group 2 (11–12)	0.71	0.85
Age Group 3 (13–14)	0.77	0.89

For Age Group 1, the conventional Cronbach's alpha (raw alpha) was 0.71, while the deflation-corrected alpha was 0.85. Although the raw alpha indicates acceptable reliability, the deflation-corrected estimate suggests good internal consistency. According to George and Mallery (2003), values above 0.80 are generally considered indicative of good reliability. Two items showed deflation-corrected item–rest correlations (DIR) below the 0.30 threshold: ALG.13.B (DIR = 0.29) and OTH.09.A (DIR = 0.28). These items were classified as underperforming and excluded from subsequent analyzes. After their removal, the deflation-corrected alpha remained stable at 0.85, and the raw alpha was 0.70.

For Age Group 2, the raw Cronbach's alpha was 0.71, and the deflation-corrected alpha was 0.85, indicating good reliability after correction. Two items demonstrated insufficient internal consistency (DIR < 0.30): ALG.03.A (DIR = 0.26) and OTH.08 (DIR = 0.11). These items were removed from further analyzes. Following their exclusion, the deflation-corrected alpha increased slightly to 0.86 and the raw alpha to 0.72.

For Age Group 3, the raw Cronbach's alpha was 0.77, and the deflation-corrected alpha was 0.89, indicating strong internal consistency. No items exhibited DIR values below 0.30 in this group.

Across all age groups, the deflation-corrected alpha coefficients exceeded 0.80, demonstrating good internal consistency of the COMATH CT assessment. Two items in Age Groups 1 and 2 were excluded due to weak item–rest correlations, while no exclusions were necessary for Age Group 3.

These results indicate that the refined COMATH assessment provides stable and internally coherent measurement across age levels. The slightly higher reliability observed in the oldest age group is consistent with improved alignment between item difficulty and student ability, as shown in the subsequent IRT analyzes. Overall, the reliability findings support the use of the instrument for comparative analyzes across age groups and countries.

5.2 Structural validity of the CT assessment

The structural validity of the COMATH CT assessment was examined using Confirmatory Factor Analysis (CFA). The CFA results indicate that across age groups, the one-factor model provided a better fit to the data than the two-factor model. This suggests that the assessment items reflect a single underlying CT construct rather than two distinct dimensions corresponding to the ALG and OTH categories. Model fit was strongest for Age Groups 2 and 3, where the one-factor model met the predefined criteria (except for the chi-square

test, which is sensitive to large sample sizes). For Age Group 1, model fit indices did not fully meet the strict thresholds, but the values (CFI ≈ 0.90; RMSEA ≈ 0.06) indicated acceptable fit. Overall, model diagnostics improved with age, suggesting better alignment between item structure and the latent construct among older students.

It should be noted that factor-analytic techniques traditionally assume continuous observed variables, whereas the COMATH items are dichotomous. Although binary responses can be interpreted as manifestations of an underlying continuous ability, this assumption may attenuate factor loadings. Therefore, the structural findings should be interpreted alongside the IRT analyzes, which are specifically designed for dichotomous data and provide complementary evidence on item functioning.

5.3 Item-Level Functioning and Measurement Precision

Item functioning was examined using a two-parameter IRT model, which estimates an item discrimination parameter (a) and an item difficulty parameter (b) for each test item. The distributions of difficulty estimates are visualized in Figures 9–11, and full parameter estimates are reported in Tables 5–7.

For Age Group 1, all discrimination estimates exceeded 0.65 (Table 5), indicating that each item contributed meaningfully to differentiating students across the ability continuum. Difficulty estimates were distributed primarily between -1 and 2 (Figure 9), with a clear tendency toward positive values. This pattern suggests that although the items span a reasonably broad range from moderately easy to difficult, the test is somewhat demanding for this age group and provides relatively little coverage of very easy items targeting lower-ability students. Overall, the parameter estimates do not indicate concerns about item behavior, and the item set demonstrates stable functioning with moderate-to-strong discrimination and a generally higher difficulty profile.

Table 5. Parameter estimates of the 2-parameter IRT model for Pilot 2 CT test (Age Group 1)

Item	Discrimination	Difficulty	Guessing
ALG.01	1.42	-1.05	0.25
ALG.02.B	2.14	1.12	0.25
ALG.08.B	1.57	-0.67	0.25
ALG.09.A	1.53	0.46	0
ALG.11.B	1.09	0.54	0
ALG.12.B	3.02	1.06	0.17
OTH.01.A	1.29	-1.1	0.2
OTH.03.B	1.92	2.1	0
OTH.08	0.9	-0.73	0.25
OTH.10.A	0.92	0.85	0
OTH.11.A	1.65	0.84	0.25
OTH.12.A	1.2	-0.08	0

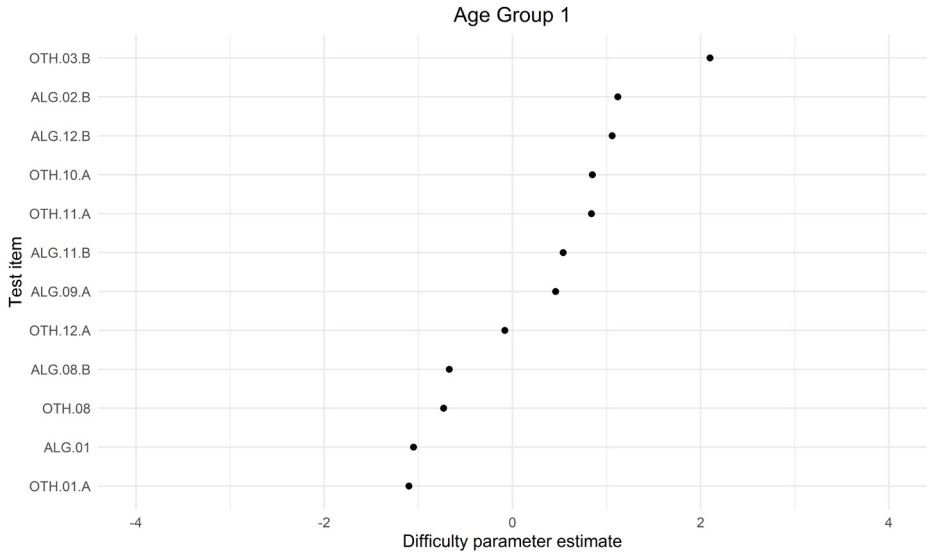


Figure 9. Distribution of the difficulty parameter estimates of the 2-parameter IRT model for Pilot 2 CT test Age Group 1.

In Age Group 2, discrimination estimates again consistently exceeded 0.65 (Table 6), supporting adequate differentiation across student ability levels. Difficulty estimates were more evenly distributed across approximately -2 to 2 (Figure 10), indicating improved targeting compared to Age Group 1. At the same time, only a small number of items fell clearly into the low-difficulty range (below approximately -0.5), suggesting that the assessment remains weighted toward moderate and more challenging items. Nevertheless, the combination of uniformly acceptable discrimination and a broad difficulty span indicates that the test items function well psychometrically for this age group.

Table 6. Parameter estimates of the 2-parameter IRT model for Pilot 2 CT test (Age Group 2)

Item	Discrimination	Difficulty	Guessing
ALG.01	1.22	-1.86	0.25
ALG.10.A	1.23	1.38	0
ALG.11.B	1.5	-0.18	0
ALG.12.B	1.75	0.31	0.17
ALG.13.B	1.65	1.02	0.1
ALG.14	1.06	-0.24	0.25
OTH.09.A	1.79	0.59	0.25
OTH.10.A	0.97	0.03	0
OTH.11.A	1.62	-0.28	0.25
OTH.12.A	1.12	-1.11	0
OTH.13.B	1.85	1.08	0
OTH.15.A	1.49	1.68	0

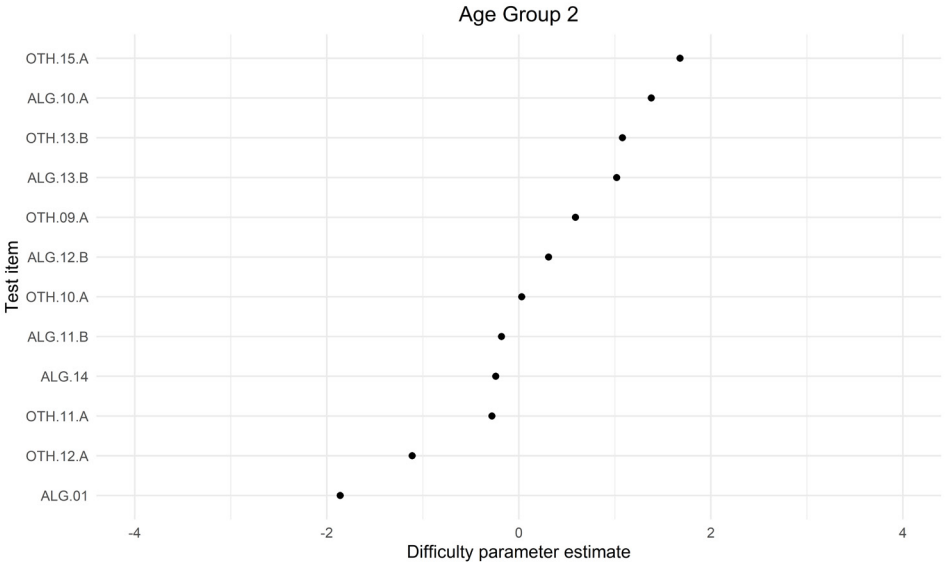


Figure 10. Distribution of the difficulty parameter estimates of the 2-parameter IRT model for Pilot 2 CT test Age Group 2.

For Age Group 3, all discrimination values exceeded 0.65 (Table 7), indicating effective differentiation. One item (ALG.04.A) showed a discrimination estimate slightly above the upper guideline (a = 4.28). As this is an isolated case and close to the cut-off, it was not treated as evidence of instability, but it was flagged for consideration in ongoing refinement. Importantly, difficulty estimates were distributed evenly within the targeted range of -2 to 2 (Figure 11), suggesting strong alignment between item difficulty and student ability in this group. Overall, Age Group 3 displayed the most balanced difficulty coverage and, correspondingly, the best targeting among the three age groups.

Table 7. Parameter estimates of the 2-parameter IRT model for Pilot 2 CT test (Age Group 3)

Item	Discrimination	Difficulty	Guessing
ALG.03.A	2.42	1.06	0.25
ALG.04.A	4.28	1.04	0.25
ALG.06.B	1.64	1.78	0
ALG.10.A	1.86	0.48	0
ALG.11.B	1.4	-0.97	0
ALG.12.B	1.25	-0.39	0.17
ALG.14	0.94	-0.76	0.25
OTH.04.B	1.66	1.4	0
OTH.09.A	1.56	-0.1	0.25
OTH.10.A	0.93	-0.71	0
OTH.11.A	1.16	-1.1	0.25
OTH.12.A	1.22	-1.73	0
OTH.13.B	1.46	0.27	0
OTH.15.A	2.02	0.77	0

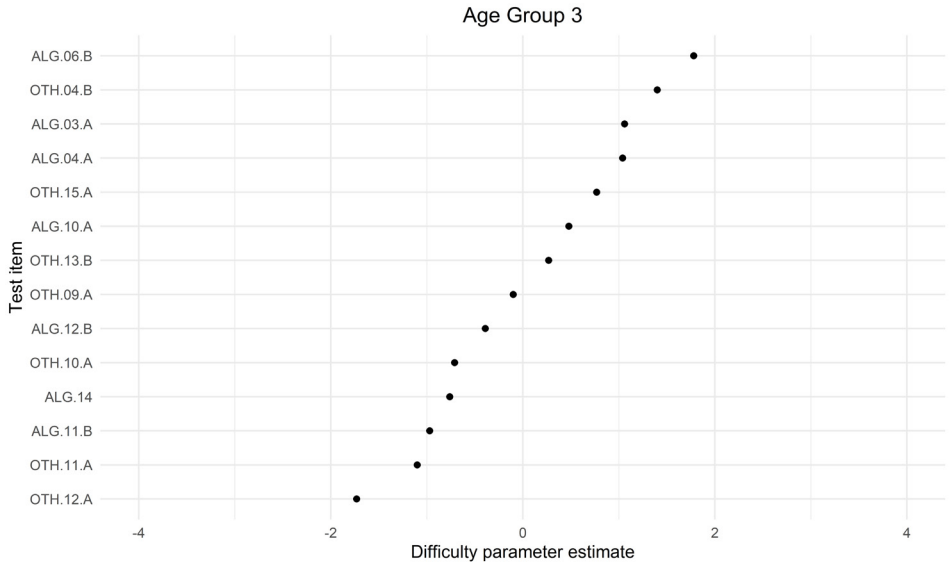


Figure 11. Distribution of the difficulty parameter estimates of the 2-parameter IRT model for Pilot 2 CT test Age Group 3.

Across all age groups, the two-parameter IRT model results indicate satisfactory item behavior: discrimination estimates were consistently at least moderate, and difficulty parameters covered a broad range of ability levels (Tables 5–7). However, the difficulty distributions in Figures 9 and 10 show that the tests for Age Groups 1 and 2 are somewhat skewed toward higher difficulty, which may reduce measurement precision for lower-performing students. In contrast, Age Group 3 shows the most even distribution of difficulty across −2 to 2, suggesting the strongest match between item difficulties and the ability range of the target population. Taken together, these results support the psychometric quality of the refined COMATH CT test in Pilot 2 and provide evidence that item selection and calibration improved overall measurement targeting, particularly for older students.

5.4 Age-Related Differences and Developmental Appropriateness

Comparisons across age groups revealed systematic differences in item functioning and test difficulty that are consistent with known developmental trajectories in cognitive and problem-solving skills. Younger students required more time to complete the assessment and were more affected by item difficulty, particularly when tasks involved higher levels of abstraction or complex reasoning. These findings highlight the importance of age-specific calibration when assessing CT in primary and lower secondary education.

Figures 12–14 present the test information curves and corresponding standard error functions for the COMATH CT-assessment instrument in the second pilot, estimated using a two-parameter IRT model for each age group. In all three figures, the solid line represents

test information, while the dashed line indicates the standard error of ability estimation, SE(θ), which is inversely related to information.

For Age Group 1, the test information curve peaks at approximately $\theta \approx 1$, indicating that the assessment provides the greatest measurement precision for students with slightly above-average CT ability. Information decreases toward both ends of the ability scale, with a corresponding increase in standard error for very low- and very high-ability students. This pattern suggests that, despite refinements introduced after the first pilot, the test remains more sensitive to higher-ability levels among younger learners, while precision is reduced for students at the lower and higher ends of the ability distribution.

For Age Group 2, the information curve shows a similar but slightly broader profile. The peak information is centred around θ values between approximately 0.5 and 1. This indicates improved measurement precision for students with medium CT ability. Although precision still decreases at the extremes, the flatter information curve suggests that the assessment is better balanced for this age group than for younger students.

For Age Group 3, the test information curve reaches a higher peak and is more sharply defined, with maximum information concentrated around $\theta \approx 1$. The information curve has higher values than in Age Groups 1 and 2 across almost the entire θ range. This indicates that the assessment is most precisely targeted for older students and provides more accurate ability estimates across a wider range of CT competence levels. While precision again decreases at the extremes of the scale, the overall shape of the curve suggests stronger alignment between item difficulty and student ability in this age group.

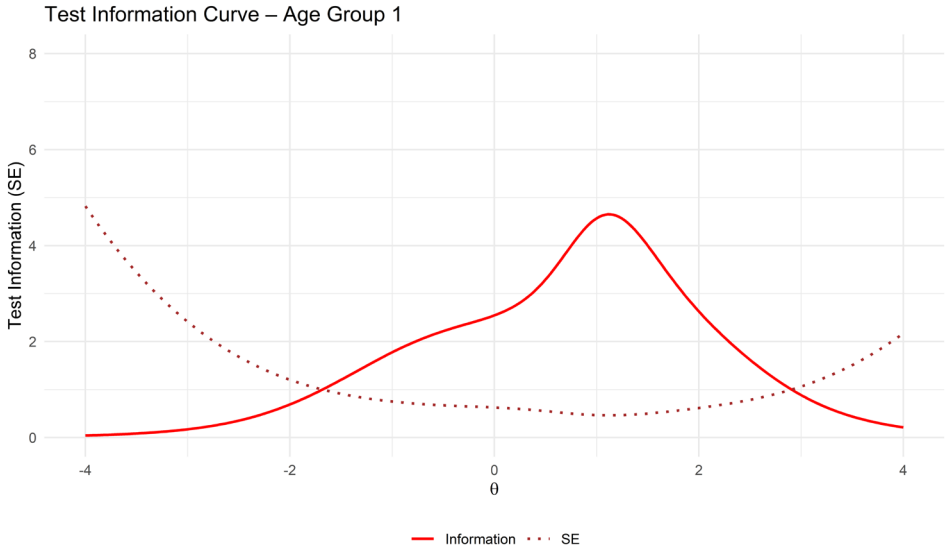


Figure 12. Test information curve of the 2-parameter IRT model for Pilot 2 CT test Age Group 1.

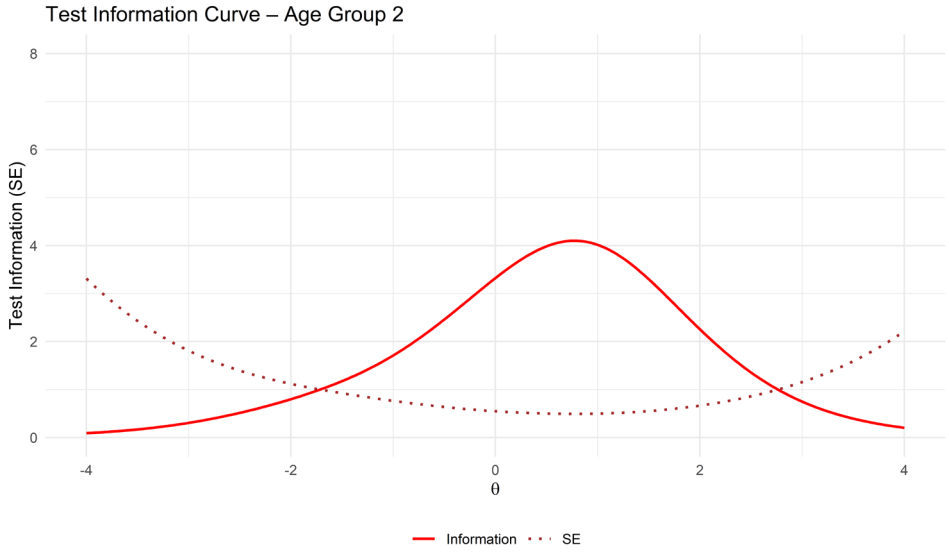


Figure 13. Test information curve of the 2-parameter IRT model for Pilot 2 CT test Age Group 2.

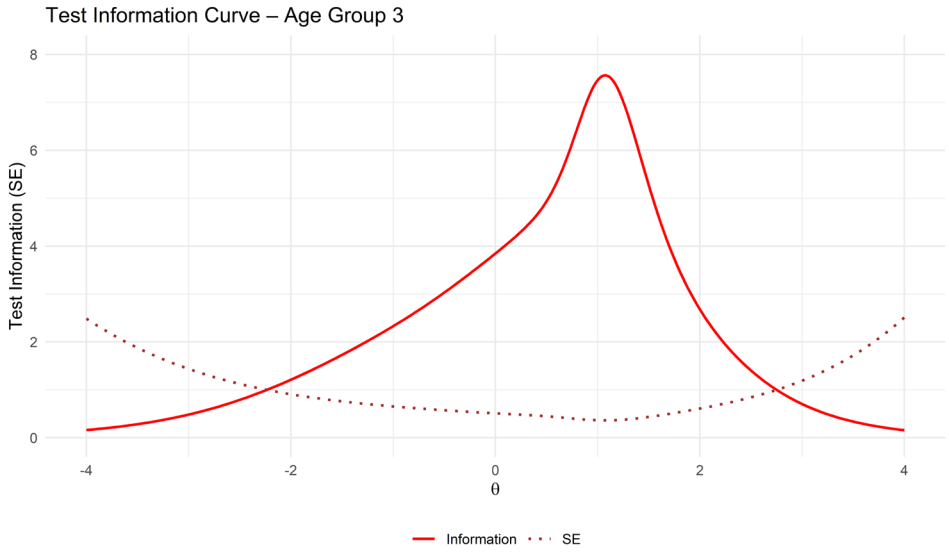


Figure 14. Test information curve of the 2-parameter IRT model for Pilot 2 CT test Age Group 3.

Across age groups, a clear developmental trend emerges. Measurement precision improves progressively from Age Group 1 to Age Group 3, as evidenced by higher information values across broader ranges of the ability scale. This pattern reflects the increasing alignment between task complexity and students' cognitive and problem-solving capacities with age.

The results also confirm that the refinements introduced after the first pilot enhanced the targeting of the assessment, particularly for older students.

At the same time, the information curves indicate that the COMATH CT-assessment is generally most precise for students with average to above-average CT ability, while precision is lower for students at the extremes of the ability distribution. This finding is typical for fixed-length assessments and highlights potential directions for future development, such as the inclusion of additional lower-difficulty items for younger learners or the adoption of adaptive testing approaches to improve precision across the full ability range.

Overall, the test information analyzes provide further evidence of the psychometric quality of the COMATH assessment and support its suitability for assessing CT across age groups, while also offering clear guidance for continued refinement and optimisation.

The revised COMATH assessment demonstrated improved alignment between item difficulty and student ability across all age groups, with particularly strong psychometric performance observed for students aged 13–14. For younger learners, the results underscore the need for careful attention to linguistic complexity, task representation, and cognitive load in CT assessment design. Importantly, these age-related differences should not be interpreted as limitations of the instrument but rather as evidence supporting the necessity of iterative, developmentally informed assessment design.

By incorporating age-specific versions and refining tasks based on empirical evidence, the COMATH framework illustrates how CT assessments can be adapted to reflect students' evolving cognitive capacities while maintaining construct coherence across age levels.

5.5 Cross-Country Comparability and Educational Applicability

Analyzes across participating countries suggest that the COMATH assessment demonstrates a reasonable level of consistency in diverse educational contexts. Although differences in student performance and item functioning were observed, these variations were primarily associated with developmental and contextual factors rather than systematic bias related to language or national curricula. The use of common item sets, anchor tasks, and uniform psychometric criteria enabled meaningful cross-national comparisons and supported the fairness and comparability of results.

Cross-country structural validity was examined using a reference-country approach. Lithuania was selected as the reference country due to its substantially larger sample size across age groups. A one-factor CFA model was first fitted to the Lithuanian data and, where necessary, adjusted to optimise model fit. The same model was then applied to data from other countries to evaluate whether the underlying structure generalised across countries. As in previous analyzes, model fit was assessed using scaled indices (Chi-square test $p > .05$; CFI > .95; TLI > .95; RMSEA < .06) with the DWLS estimator.

Because some country- and age-group-specific samples were relatively small, CFA was conducted only for countries with at least 100 participants. This included Spanish data for Age Group 2 and all countries except Türkiye for Age Group 3.

For Age Group 2, the one-factor model (after excluding item OTH.10.A) demonstrated good fit for the Lithuanian data (CFI = 0.972; TLI = 0.965; RMSEA = 0.037), although the chi-square test was significant, likely due to the large sample size (n = 715). When applied to the Spanish Age Group 2 data, the model showed excellent fit across all indices, indicating strong structural consistency across these two countries.

For Age Group 3, the one-factor model including all 14 items provided a good fit for the Lithuanian data (CFI = 0.984; TLI = 0.981; RMSEA = 0.030), again with a significant chi-square value attributable to sample size (n = 556). When the same model was applied to other countries, results were mixed. The Finnish data demonstrated good model fit across all indices, supporting structural generalisability. The Spanish data showed acceptable fit, with CFI and TLI values slightly below the recommended threshold but sufficiently close to be considered satisfactory. In contrast, the Hungarian and Swedish datasets did not meet the predefined fit criteria, and in the Hungarian case, the model produced an inadmissible solution (negative latent variance), suggesting instability. These discrepancies may reflect contextual differences or limited sample sizes rather than fundamental structural divergence.

Summarizing the results, the COMATH CT assessment shows promising cross-national applicability, particularly in contexts with sufficient sample size and stable data conditions. At the same time, the findings underscore the importance of contextual sensitivity and adequate sampling when evaluating structural equivalence across countries. Together, these results support the potential of COMATH as a scalable assessment instrument while emphasizing the need for continued cross-national validation.

5.6 Synthesis of Findings

Taken together, the results show that the COMATH assessment instrument provides a valid, reliable, and developmentally appropriate measure of CT skills for students aged 9–14. Factor analyzes support the interpretation of CT as a unified competence within the assessment, while IRT analyzes highlight the critical role of item-level evaluation in achieving balanced difficulty and strong discrimination. Age-group comparisons and cross-national analyzes further demonstrate the importance of iterative refinement and contextual sensitivity in CT assessment design.

Importantly, the findings illustrate that CT assessment should not be viewed as a static measurement exercise but as a dynamic, iterative process that links task design, empirical validation, and instructional use. This perspective aligns with contemporary approaches to assessment for learning and supports the integration of CT assessment into everyday classroom practice.

The findings of this study provide important insights into the assessment of CT in school education and contribute to ongoing discussions about how complex cognitive competences can be measured in developmentally appropriate and psychometrically robust ways. By combining factor-analytic methods, IRT, and cross-national evidence, the COMATH as-

sessment instrument offers a comprehensive perspective on both the theoretical and practical challenges of CT assessment.

First, the factor-analytic results support the interpretation of CT as a largely unified latent construct rather than a set of clearly separable subskills. Although the COMATH assessment was designed to include tasks targeting different CT components, CFA did not identify distinct conceptual dimensions aligned with these components. Instead, CFA supported a 1-factor structure in which all items reflect a single underlying CT construct. From an assessment perspective, this supports the use of coherent, integrated task sets rather than overly fine-grained subscales that may not be empirically distinguishable at this developmental stage.

Second, the IRT results highlight the critical importance of item-level analysis for achieving measurement precision and fairness in CT assessment. The initial skew in item difficulty toward higher ability levels, especially for younger students, demonstrates that even theoretically well-designed tasks can misalign with learners' developmental capacities. The improvements observed following IRT-informed refinement confirm that iterative, data-driven revision is essential for optimising assessment quality. These findings reinforce the value of IRT not only as an analytical tool but also as a design instrument that supports principled decision-making about item selection, revision, and test length.

Third, age-related differences observed across the analyzes underscore the necessity of developmentally sensitive assessment design. While the revised COMATH CT-assessment instrument demonstrated strong psychometric performance for students aged 13–14, younger learners were more affected by item difficulty, cognitive load, and linguistic demands. Importantly, these differences should not be interpreted as shortcomings of the assessment but rather as evidence of the complex interaction between task design and cognitive development. The use of age-specific versions within a common assessment framework proved effective in maintaining construct coherence while allowing for appropriate differentiation, illustrating a viable approach for CT assessment across broad age ranges.

From a cross-national perspective, the results provide encouraging evidence of the comparability and fairness of the COMATH CT-assessment. Despite differences in educational systems, curricula, and classroom practices, the assessment functioned consistently across participating countries. Observed variations in performance were largely attributable to contextual and developmental factors rather than systematic bias, supporting the applicability of COMATH as a cross-national CT assessment instrument. This finding is particularly relevant in light of growing international interest in benchmarking CT skills and developing shared assessment frameworks.

Beyond psychometric considerations, integrating teacher and student feedback underscores the importance of practical validity in CT assessment. Teachers' reports of feasibility and curricular alignment, together with students' perceptions of task clarity and relevance, indicate that COMATH supports assessment not only of learning but also for learning. This dual role is especially important for CT, where assessment outcomes can inform instructional planning, differentiation, and targeted support.

Taken together, the findings suggest that effective CT assessment requires a balanced integration of theoretical grounding, psychometric rigour, developmental sensitivity, and pedagogical relevance. The COMATH framework demonstrates how these elements can be combined through an iterative design and validation process. More broadly, the results reinforce the view that CT assessment should be understood as a dynamic process that evolves alongside learners, instructional practices, and educational contexts, rather than as a static measurement exercise.

6. Conclusions

This study examined the design, validation, and cross-national applicability of the COMATH CT assessment instrument for students aged 9–14 (Grades 3–8). Through a systematic development process combining theoretical grounding, expert evaluation, iterative pilot studies, and advanced psychometric modelling, the study provides empirical evidence addressing the three research questions and contributes to the growing body of research on CT assessment in school education.

Regarding RQ1, the findings demonstrate that age-appropriate CT assessment tasks can be systematically designed by integrating conceptual frameworks, large-scale task analytics, expert-based content validation, and empirical calibration. The combination of Bebras-inspired problem formats, explicit alignment with CT components, and developmental calibration across three age groups enabled the COMATH assessment to reflect differences in students' cognitive maturity and problem-solving capacities. The results further highlight the importance of balancing task difficulty, linguistic clarity, and cognitive load, particularly for younger learners, in order to ensure both measurement validity and classroom feasibility.

Regarding RQ2, psychometric analyzes provide consistent evidence of acceptable structural validity, reliability, and item-level functioning across age groups. Factor-analytic results support interpreting CT as a largely unified latent construct within the COMATH framework, with test items reflecting a single CT factor rather than sharply separable subskills. IRT analyzes confirm that, following refinement, the instrument contains items with appropriate discrimination and a balanced range of difficulty levels. The combined use of CFA and IRT demonstrates the methodological robustness of the validation process and illustrates how item-level diagnostics can meaningfully inform iterative assessment design.

Regarding RQ3, cross-national analyzes indicate that the COMATH CT assessment shows reasonable structural consistency across participating countries. While some variability in model fit emerged, particularly in smaller national samples, no systematic evidence of linguistic or curricular bias was identified. The use of common anchor items and harmonised psychometric criteria enabled meaningful cross-country comparisons. Together with positive teacher and student evaluations, these findings support the applicability of COMATH across diverse educational contexts and underscore the importance of adequate sampling in cross-national validation.

Importantly, COMATH extends existing CT assessment approaches in several ways. Un-

like instruments such as the CTDI/CTt (Román-González et al., 2017; Román-González et al., 2019; Zapata-Cáceres et al., 2020), TechCheck (Relkin et al., 2020), CTA-CES (Li et al., 2021), or Bebras Cards (Dagienė & Sentance, 2016), which are typically designed for a single age range, specific learning environment, or limited validation context, COMATH was explicitly developed as:

- a vertically aligned assessment spanning three developmental stages (9–14 years),
- a cross-nationally piloted instrument with harmonized psychometric validation,
- a task set grounded in large-scale item analytics from international Bebras data,
- and a framework integrating rigorous psychometric modelling and pedagogical usability.

Rather than focusing solely on summative measurement, COMATH was designed to balance measurement precision with classroom applicability, supporting both comparative research and instructional reflection. In this sense, the study contributes not only a validated instrument but also a replicable development model for CT assessment.

Overall, the findings reinforce the view that effective CT assessment requires integrating psychometric rigor, developmental sensitivity, and educational relevance. The iterative refinement process adopted in COMATH illustrates how empirical evidence can directly inform task selection, calibration, and optimization.

Future research will focus on examining longitudinal stability, expanding cross-national validation to additional contexts, and exploring adaptive or formative extensions of the instrument.

The complete set of COMATH CT tasks, finalised after the second pilot, is provided in Appendix 1.

Acknowledgments

This work was co-funded by the European Union. The study was conducted with the framework of the CT&MathABLE project, coordinated by Vilnius University. Further information about the project is available at: https://www.fsf.vu.lt/en/ct-math-able. The authors sincerely thank the project partners for their collaboration and contributions: Ankara University (Türkiye), Eötvös Loránd University (Hungary), Gedminų Progymnasium (Lithuania), KTH Royal Institute of Technology (Sweden), Özkent Akbilek Middle School (Türkiye), University of the Basque Country (Spain), and the University of Turku (Finland).

The authors would also like to explicitly acknowledge and thank the members of the international Bebras challenge community who contributed to the development of the Bebras tasks used and adapted in this study. Their work substantially informed and influenced the task design process underlying the COMATH assessment.

Credits to the original authors of the Bebras tasks: Sarah Chan (Canada), Maiko Shimabuku (Japan), Valentina Dagienė (Lithuania), Anupama Sivakumar (India), Chris Roffey (UK), Taina Lehtimäki & Tom Naughton (Ireland), Leonardo Cavalcante (Brazil), Byeonggyu Cho, Seulki Kim, Jihye Kim & Hakin Kim (South Korea), Marielle Léonard (France), Bernadette Spieler, Tobias Berner & Susanne Datzko (Switzerland), Marious o. Choudary (Pakistan), Le Quang Quan (Vietnam), Michael Weigend (Germany), Mark Edward M. Gonzales (Philippines), Thomas Ioannou (Cyprus), Adam Grodec & Susannah Quidilla (Australia), Troy Vasiga (Canada), Jiří Vaníček (Czechia), Yasemin Gulbahar (Turkey), Zsuzsa Pluhár (Hungary). The authors also acknowledge Vaidotas Kinčius, designer of the Bebras task illustrations. Where applicable, adapted tasks were modified for age appropriateness and alignment with the COMATH assessment framework in accordance with the CC BY-SA 4.0 licence requirements.

Statements of open data and ethics

This study was conducted across multiple European countries in regular school settings using the ViLLE digital learning environment, which participating schools routinely use for instructional purposes. The pilot assessments formed part of normal classroom activities in which students solved learning tasks within the platform.

The study involved minor participants (students aged 9–14). No sensitive personal data was collected. Only students' task responses and minimal background variables necessary for analysis (e.g., age group and gender, where provided) were recorded. No identifying information, such as names or personal identification numbers, was included in the research dataset. All data were anonymised prior to analysis.

Informed consent procedures were managed locally by participating schools in accordance with national regulations. Schools confirmed that students and their legal guardians had been informed about the research and had provided the necessary permissions for participation in research activities conducted within the learning environment. Participation was voluntary, and students could withdraw from data use at any stage prior to analysis.

The research complied with the General Data Protection Regulation (GDPR). All collected data were securely stored on protected servers at the University of Turku, with access restricted to authorised members of the research team.

Due to the nature of this research study, no separate institutional review board or research ethics committee approval number was required under the applicable national regulations of the participating institutions.

Artificial intelligence tools (ChatGPT) were used solely for language editing and grammatical refinement of the manuscript. No AI tools were used in the study design, data collection, statistical analysis, or interpretation of results.

References

Almanasreh, E., Moles, R. & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. Research in Social and Administrative Pharmacy, 15(2), 214–221. https://doi.org/10.1016/j.sapharm.2018.03.066

Al-Shehri, M. A. M. (2020). The effectiveness of the collaborative investigation strategy in the achievement and development of algebraic thinking skills for first intermediate grade students. Journal of the College of Basic Education for Educational and Human Sciences, 46(1), 259–272.

Araujo, A. L. S. O., Andrade, W. L., Guerrero, D. D. S. & Melo, M. R. A. (2019). How Many Abilities Can We Measure in Computational Thinking?: A Study on Bebras Challenge. Proceedings of the 50th ACM Technical Symposium on Computer Science Education, 545–551. https://doi.org/10.1145/3287324.3287405

Armoni, M. (2016). Computer science, computational thinking, programming, coding: the anomalies of transitivity in K-12 computer science education. In ACM Inroads, 7(4), 24–27. https://doi.org/10.1145/3011071

Baker, F. B. (2001). The basics of item response theory (2nd. Ed.). Retrieved from https://eric.ed.gov/?id=ED458219.

Basu, S., Rutstein, D. W., Xu, Y., Wang, H. & Shear, L. (2023). A principled approach to designing computational thinking concepts and practices assessments for upper elementary grades. In Assessing Computational Thinking (pp. 57–86). Routledge.

Bentler, P. M. & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88(3), 588–606. https://doi.org/10.1037/0033-2909.88.3.588

Bilbao, J., Bravo, E., García, O., Rebollar, C., Feniser, C., Dagienė, V., Masiulionytė-Dagienė, V., Laakso, M.-J., Hakkarainen, A. Güven, I., Gulbahar, Y., Pluhár, Z., Sarmasági, P. & Pears A. (2023). Algebraic Thinking and Computational Thinking in Pre-University Curriculum, INTED2023 Proceedings, 3888–3895.

Bilbao, J., Bravo, E., Garcia, O., Rebollar, C., Laakso, M. J. Kaarto, H., Lehtonen, D., Parviainen, M., Jankauskiene, A., Pears, A., Guven, I., Gülbahar, Y., Öztürk, H. T., Tan Yenigün, N., Pluhár, Z., Sarmasági, P., Rumbus, A., Dagienė, V. & Masiulionytė-Dagienė, V. (2025). Analytical Methods and Tools for Evaluating the Development of Computational Thinking Abilities", International Journal of Education and Information Technologies, 19, pp. 53–61.

Bocconi, S., Chioccariello, A., Kampylis, P., Dagienė, V., Wastiau, P., Engelhardt, K., Earp, J., Horvath, M.A., Jasutė, E., Malagoli, C., Masiulionytė-Dagienė, V., Stupurienė, G. (2022). Reviewing Computational Thinking in Compulsory Education. Publications Office of the European Union. https://doi.org/10.2760/126955

Chen, G., Shen, J., Barth-Cohen, L., Jiang, S., Huang, X. & Eltoukhy, M. (2017). Assessing elementary students' computational thinking in everyday reasoning and robotics programming, Computers & education, 109, 162–175.

Dagiene, V., Hromkovic, J., Lacher, R. (2021). Designing informatics curriculum for K-12 education: From Concepts to Implementations. Informatics in Education, 20, 3 (September 2021), 333–360. https://doi:10.15388/infedu.2021.22

Dagienė, V., Kampylis, P., Giannoutsou, N., Engelhardt, K., Malagoli, C. & Bocconi, S. (2024). Fostering computational thinking in compulsory education in Europe: a multiple case study. Baltic

journal of modern computing, 12(2), 189–221. doi:10.22364/bjmc.2024.12.2.05

Dagienė, V. & Sentance, S. (2016). It's computational thinking! Bebras tasks in the curriculum. In International conference on informatics in schools: Situation, evolution, and perspectives (pp. 28–39). Cham: Springer International Publishing.

Dolgopolovas, V. & Dagienė, V. (2024). Competency-based TPACK approaches to computational thinking and integrated STEM: A conceptual exploration. Computer applications in engineering education, 32(6), 1–28. doi:10.1002/cae.22788

Erola, K. & Mirel, J. (2023). Ohjelmoinnillisen ajattelun mittarit: Systemaattinen kirjallisuuskatsaus [Computational thinking assessment instruments: A systematic literature review, Master's thesis, University of Turku]. UTUPUB. https://www.utupub.fi/handle/10024/175021

Ezeamuzie, N. O. & Leung, J. S. (2022). Computational thinking through an empirical lens: A systematic review of literature. Journal of Educational Computing Research, 60(2), 481–511.

Fink, A. (2019). Conducting research literature reviews: From the internet to paper. Sage publications.

Gane, B. D., Israel, M., Elagha, N., Yan, W., Luo, F. & Pellegrino, J. W. (2021). Design and validation of learning trajectory-based assessments for computational thinking in upper elementary grades. Computer Science Education, 31(2), 141–168.

George, D. & Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.). Boston: Allyn & Bacon.

Gülbahar, Y., Kert, S. B., Kalelioğlu, F. (2019). The self-efficacy perception scale for computational thinking skill: Validity and reliability study, Turkish Journal of Computer and Mathematics Education, 10(1), 1–29.

Gulbahar, Y., Öztürk, T., Dagienė, V., Parviainen, M., Güven, I. & Bilbao, J. (2025). Evaluating Interactive Tasks through the Lens of Computational and Algebraic Thinking, Interactivity Types, and Multimedia Design Principles, Olympiads in Informatics, Vol. 19, 63–86, DOI:10.15388/ioi.2025.05 2025.

Gyamfi, A. & Acquaye, R. (2023). Parameters and Models of Item Response Theory (IRT): A Review of Literature", Acta Educationis Generalis, 13(3), pp. 68–78, https://doi.org/10.2478/atd-2023-0022[3] (https://sciendo.com/article/10.2478/atd-2023-0022)

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, (November 2018), 296–310. https://doi.org/10.1016/j.compedu.2018.07.004

Kalelioglu, F., Gülbahar, Y. & Kukul, V. (2016). A framework for computational thinking based on a systematic research review", Baltic Journal of Modern Computing, 4(3), 583.

Kampylis, P., Dagienė, V., Bocconi, S., Chioccariello, A., Engelhardt, K., Stupurienė, G., Masiulionytė-Dagienė, V., Jasutė, E., Malagoli, C., Horvath, M. & Earp, J. (2023). Integrating Computational Thinking into Primary and Lower Secondary Education: A Systematic Review. Educational Technology & Society, 26(2), 99–117. https://doi.org/10.30191/ETS.202304_26(2).0008

Kong, S.C. & Wang, Y.Q. (2021). Item response analysis of computational thinking practices: Test characteristics and students' learning abilities in visual programming contexts", Computers in Human Behavior, 122, 106836.

Laakso M., Kaila, E. & Rajala, T. (2018). ViLLE–collaborative education tool: Designing and utilizing an exercise-based learning environment. Education and Information Technologies, 23, 1655–167

Lawshe, C. H. (1975). A quantitative approach to content validity. Personnel Psychology, 28(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lehtonen, D., Satoma, E., Kaarto, H., Parvioainen, M., Laakso, M.-J., Sarmasági, P., Pluhár, Z., Rumbus, A., Dagienė, V., Masiulionytė-Dagienė, V., Jankauskiene, A., Bilbao, J., Bravo, E., García, O., Rebollar, C., Pears, A., Güven, I., Gulbahar, Y., Öztürk, T. & Yenigün, N.T. (2025). Developing and evaluating an online assessment of computational and algebraic thinking: perspectives of students and teachers from six countries. Proceedings of the 17th annual International Conference on Education and New Learning Technologies (EDULEARN25), Palma, Spain (pp. 4570–4578). https://doi.org/10.21125/edulearn.2025.1183

Li, Y., Xu, S. & Liu, J. (2021). Development and validation of computational thinking assessment of Chinese elementary school students. Journal of Pacific Rim Psychology, 15, 18344909211010240.

McMillan, J. H. & Hellsten, L. (2010). Classroom assessment: Principles and practice for effective standards-based instruction. Pearson Education Canada.

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

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(4), 482–498.

Román-González, M., Moreno-León, J. & Robles, G. (2019). Combining assessment tools for a comprehensive evaluation of computational thinking interventions. In Computational thinking education (pp. 79–98). Singapore: Springer Singapore.

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.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. Journal of statistical software, 48, 1–36. https://doi.org/10.18637/jss.v048.i02

Shell, D. F. & Soh, L. K. (2013). Profiles of motivated self-regulation in college computer science courses: Differences in major versus required non-major courses. Journal of Science Education and Technology, 22(6), 899–913.

Shin, N., Bowers, J., Roderick, S., McIntyre, C., Stephens, A. L., Eidin, E., ... & Damelin, D. (2022). A framework for supporting systems thinking and computational thinking through constructing models. Instructional Science, 50(6), 933–960.

Shute, V. J., Sun, C. & Asbell-Clarke, J. (2017). Demystifying computational thinking. Educational Research Review, 22, 142–158.

Tang, X., Yin, Y., Lin, Q., Hadad, R., Zhai, X. (2020). Assessing Computational Thinking: A Systematic Review of Empirical Studies, Computers & Education, 148, 103798, 2020. https://doi.org/10.1016/j.compedu.2019.103798

Tsarava, K., Moeller, K., Román-González, M., Golle, J., Leifheit, L., Butz, M. V. & Ninaus, M. (2022). A cognitive definition of computational thinking in primary education. Computers & Education, 179, 104425.

Zapata-Cáceres, M., Martín-Barroso, E. & Román-González, M. (2021). Collaborative game-based environment and assessment tool for learning computational thinking in primary school: a case study. IEEE Transactions on Learning Technologies, 14(5), 576–589.

Zapata-Cáceres, M., Martín-Barroso, E. & Román-González, M. (2020, April). Computational thinking test for beginners: Design and content validation. In 2020 IEEE global engineering education conference (EDUCON) (pp. 1905–1914). IEEE.

Zhong, B., Wang, Q., Chen, J. & Li, Y. (2016). An exploration of three-dimensional integrated assessment for computational thinking. Journal of Educational Computing Research, 53(4), 562–590.

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L. & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms, Journal of Science Education and Technology, 25(1), 127–147. https://doi.org/10.1007/s10956-015-9581-5

Appendix 1. All the CT test items included in COMATH 1–3

Task ID (ID in the Bebras svn)	Group 1	Group 2	Group 3	Answers
F OTH-01-A (2020-CA-06, modified)	Item 1			4th teddy
F ALG-01 (2022-JP-02)	Item 2	Item 1		2nd symbol
F OTH-08-B-P2 (2022-LT-08, modified)	Item 3			2nd variant
F ALG-08-B (2022-IN-01, modified)	Item 4			the last one
F OTH-12-A (2022-UK-02)	Item 5	Item 2	Item 1	
F ALG-09-A (2022-IE-02)	Item 6			
F ALG-11-B (2022-BR-01, modified)	Item 7	Item 5	Item 3	
F OTH-11-A (2022-KR-03)	Item 8	Item 3	Item 2	2nd hamburger
F OTH-10-A (2022-FR-02)	Item 9	Item 6	Item 5	
F ALG-12-B (2022-CH-14)	Item 10	Item 7	Item 6	
F ALG-02-B (2022-PK-01, modified)	Item 11			3rd from the top
F OTH-03-B P2 (2022-VN-05, modified)	Item 12			
F ALG-14 (2022-HU-02)		Item 4	Item 4	2nd candy bag
F OTH-09-A (2022-DE-02)		Item 8	Item 7	3rd variant
F ALG-13-B-P2 (2022-PH-03, modified)		Item 9		
F OTH-13-B-P2 (F OTH-13-B, modified)		Item 10		
F ALG-10-A (2022-AU-03)		Item 11	Item 9	
F OTH-15-A (2022-SK-04)		Item 12	Item 10	HAUS
F OTH-13-B (2022-CY-01, modified)			Item 8	4-Mary, 5-Pan, 6-Niki, 7-Zac
F ALG-04-A (2022-CA-06)			Item 11	4th variant
F ALG-03-A (2022-CZ-04)			Item 12	1st plan
F OTH-04-B (2022-HU-04, modified)			Item 13	
F ALG-06-B (2022-TR-02, modified)			Item 14	

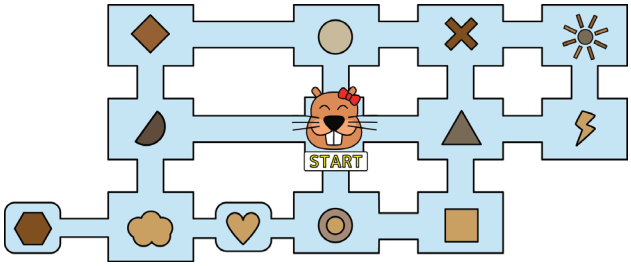
Group 1

Item 1: F OTH-01-A

Ravi wants to buy a teddy bear with a star on its foot, wearing a scarf or a bow, but not glasses. Which bear does Ravi choose?

Item 2: F ALG-01

Beaver goes from the START room to her sister's room using a map of the rooms as a guide. On the map, each room is marked by a symbol.



Beaver moves by the following arrow sequence:

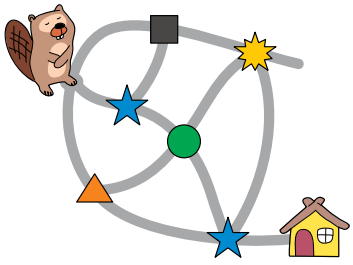
↑	→	↓	↓	↑	←	←	↓	→	→

Each arrow tells Beaver in which direction to move from one room to the next.

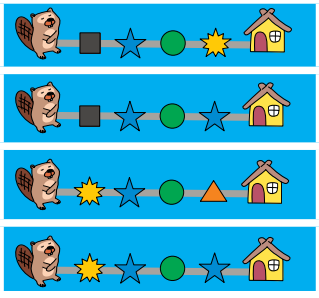
Item 3: F OTH-08-B-P2

Beaver is going home.

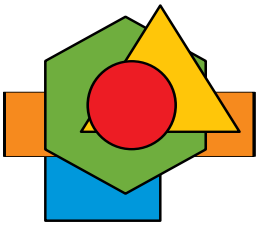
There are several different trails.



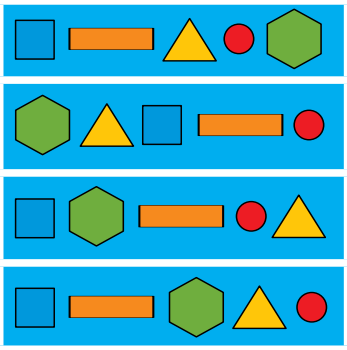
Which of the trails is correct?



Item 4: F ALG-08-B

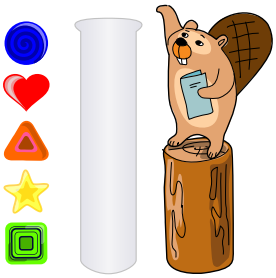


In what order have the stickers been applied?



Item 5: F OTH-12-A

Brian wants to eat five candies in the order: grape, orange, lemon, strawberry and blueberry. Only the candy on top can be eaten from the tube. Drag and drop the candies into the tube so that Brian can eat them in his preferred order.



Item 6: F ALG-09-A

The table shows which foods Betty Beaver, Fiona Fox, and Bobby Bear can eat.

	Leaves	Fish	Mushrooms	Berries
Betty Beaver	Yes	No	No	Yes
Fiona Fox	No	Yes	No	Yes
Bobby Bear	No	Yes	Yes	Yes

One day, they have nine portions of food.

Divide the food so that each of them gets three portions by dragging and dropping it onto the table.

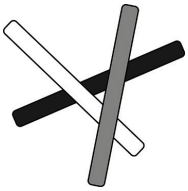
 Betty Beaver				
 Fiona Fox				
 Bobby Bear				

Item 7: F ALG-11-B

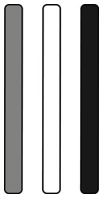
The sticks need to be picked up from a pile according to two rules:

- only pick up a stick if no other stick is covering it,
- pick up one stick at a time.

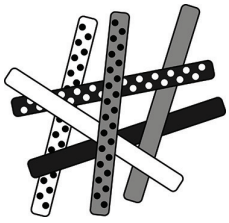
For example, if 3 sticks are in a pile like this:

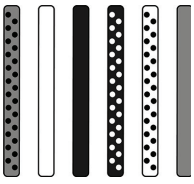


Sticks need to be picked up in this order (from left to right):



In which order should these sticks be picked up?
Drag and drop the sticks to the rectangles in correct order (from left to right).





Item 8: F OTH-11-A

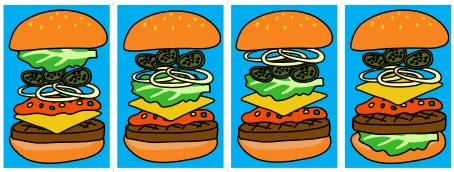
Beaver Jessica is making hamburgers according to the rules below:

Buns	Meat	Source	Pickles	Lettuce	Onions	Cheese
						

Which hamburger is correctly made according to the rules?

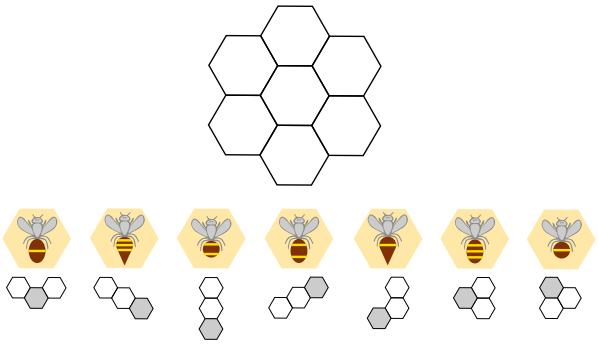
1. The sauce should be right above the meat.
2. Meat and cheese should be below the pickles, lettuce and onions.
3. Onions should not be in contact with buns.
4. All ingredients must be between the buns.

Which hamburger is correctly made according to the rules?



Item 9: F OTH-10-A

Seven bees need to fit into this hive.



Each bee has a rule: the bee must be placed in the grey cell.

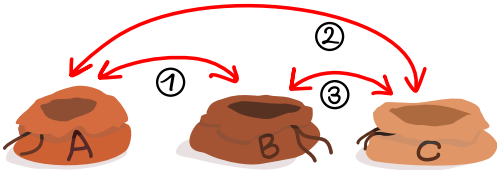
Where do each bee fit in?

Drag and drop the bees into the hive.

Item 10: F ALG-12-B

To start, the crumbled paper is put in bag A, the marble in bag B and the gem in bag C.

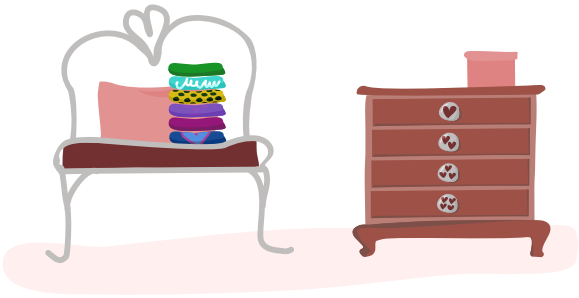
Then the items are mixed. First, items in bags A and B are switched. Then, items in bags A and C are switched. Lastly, items in bags B and C are switched.



Where are the items now? Drag and drop them into the bags.

Item 11: F ALG-02-B

There are seven shirts in a pile on the chair. Puffy puts the shirts into the drawers one by one. She starts with the top drawer, puts the first shirt in it, then puts the next shirt in the second drawer from the top, and so on. When she has put a shirt into the bottom drawer, she starts from the top again.

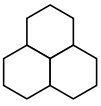


Into which drawer will she put the last shirt?

Click on the correct drawer.

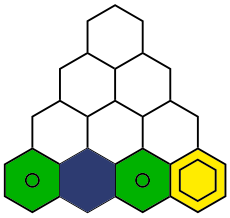
Item 12: F OTH-03-B P2

There are three different colors of hexagons. When three hexagons touch as shown, they must all be the same color or all different colors.



What does a tower like this look like when following this rule?

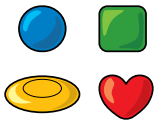
Click the hexagons multiple times to change their colors.



Group 2

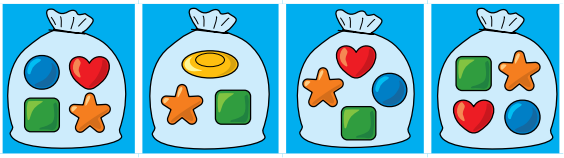
Item 4: F ALG-14

These four candies are Johnny's favorites:



If the candy bag consists of at least 2 of these, Jonny buys it.

Which of these candy bags does Jonny buy?



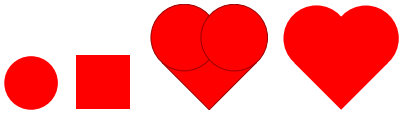
Item 8: F OTH-09-A

Tina starts with a circle and a square. She transforms these into a heart. To do this, she can only use three transformations:

Rotate the shape arbitrarily.

Move the shape arbitrarily.

Duplicate the shape in place.

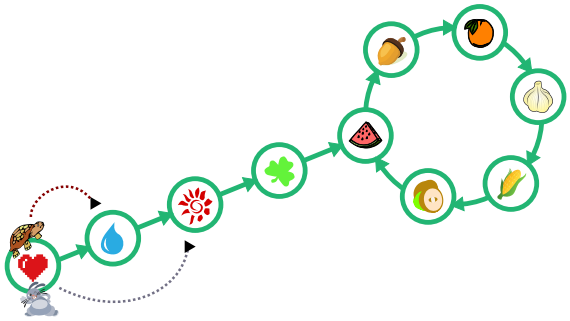


What did she do and in what order?

- Move circle, move circle, duplicate circle, move square
- Duplicate circle, rotate circle, move circle, move square
- Duplicate square, rotate square, move square, move circle
- Duplicate circle, rotate square, move circle, move circle

Item 9: F ALG-13-B-P2

A tortoise and a hare are racing against each other on a track.



They start at the same time at the circle marked with a heart symbol. They follow the arrows on the track.

In one turn, the tortoise is able to move onto the very next circle and the hare is able to move onto the second circle, skipping one circle.

In which cycle do the tortoise and the hare meet for the first time after the start? Click that circle.

Item 10: F OTH-13-B-P2

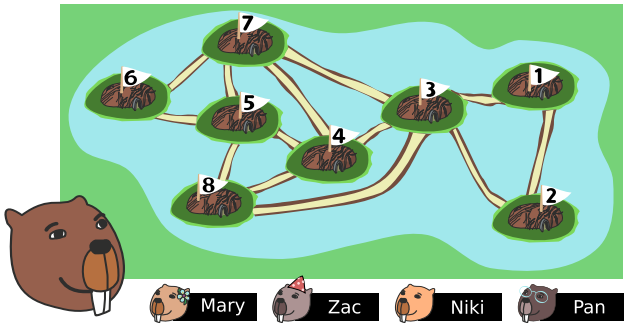
Seven beavers live in this lake.

Two beavers are neighbours if a path connects their homes. The map shows the homes and paths.

- Niki has two neighbours: Zac and Pan.
- Mary, Zac, and Pan have four neighbours each.

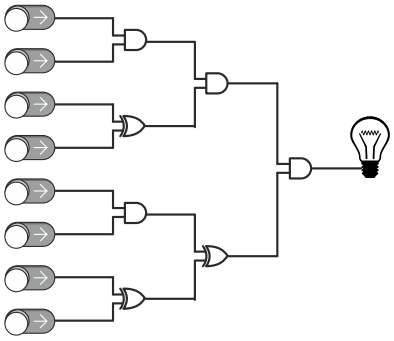
Where do beavers live?

Drag and drop the names onto the map.



Item 11: F ALG-10-A

The game “Light on” has 8 switches that can be operated. Wires lead out of these switches, then through some components, and finally to a light bulb.



The component’s output  is ON when BOTH incoming wires are ON.

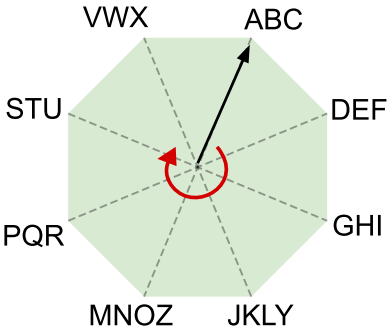
The component’s output  is ON when exactly ONE of the incoming wires is ON.

Which switches have to be ON for the light bulb to be on?

Click the switches to turn them on.

Item 12: F OTH-15-A

With this wheel, plain text is encrypted into ciphertexts:



At the start, the pointer of the wheel is set to “ABC”.

Each letter is encrypted individually, continuing where the wheel left off from the previous encryption. Two digits are determined for this purpose:

- The first digit indicates by how many positions the pointer is turned clockwise. Then the pointer is positioned on the block containing the letter to be encrypted.
- The second digit indicates the number of letters in the block to be encrypted.

For example, the word “PAAR” is encrypted as 51-31-81-53.

What does the ciphertext 22-61-62-74 mean?

Group 3

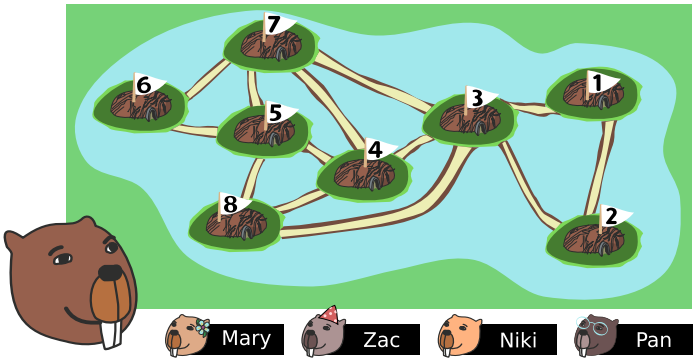
Item 8: F OTH-13-B

Eight beavers live in this lake.

Two beavers are neighbours if a path connects their homes. The map shows the homes and paths.

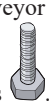
- Niki has two neighbours: Zac and Pan.
- Mary, Zac, and Pan have four neighbours each.

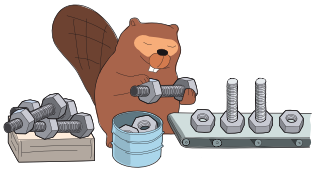
Where do beavers live? Drag and drop the names onto the map.



Item 11: F ALG-04-A

At the Beaver Construction factory, Ben works on the nuts and bolts assembly line. His job description is as follows:

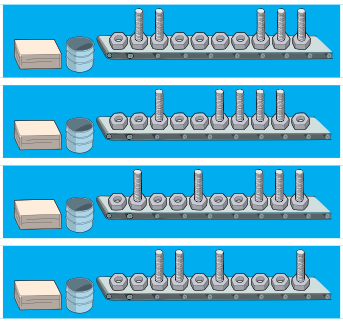
- Ben stands at one end of a long conveyor belt, which contains a line of nuts  and bolts  .
- Ben's job is to take each element, either a nut or a bolt, from the conveyor belt.
- If Ben takes a nut from the conveyor belt, he puts it in the bucket beside him.
- If Ben takes a bolt from the conveyor belt, he grabs a nut from the bucket beside him, attaches the nut and bolt together, and places the assembled part onto a large box.



However, things can go wrong for Ben in two different ways:

1. If Ben takes a bolt from the conveyor belt, and there is no nut in the bucket to attach.
2. If there are no more nuts or bolts on the conveyor belt, and there are still nuts in the bucket.

Which of these sequences of elements will not cause things to go wrong?



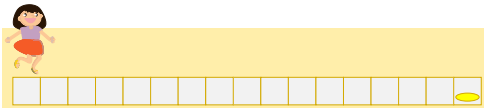
Item 12: F ALG-03-A

Veronika found 17 tiles in a line and made a game plan from them.

She put a coin at one end of a line and then stood at the other end, facing the coin (see the picture).

She wants to jump to every tile in a line using the following rules:

- If you are standing on a tile marked X, jump 3 tiles forward.
- If you are standing on a tile marked O, jump 1 tile backward.

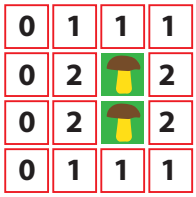


Which of the game plans will bring her to the coin?

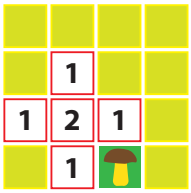
Item 13: F OTH-04-B

One toadstool is visible at the beginning of the game: “Watch out for toadstools”. All other squares on the board are covered. When you click a square, either another toadstool or the number of toadstools on the neighbouring squares appears. If you uncover all the squares without any toadstools, you win.

Here is an example of a completely uncovered board:

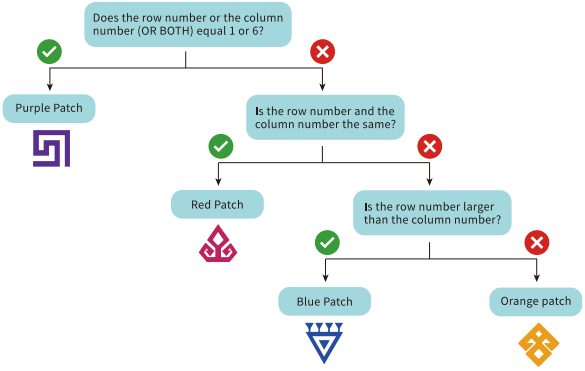


You have started a new game. In which squares is there certainly no toadstool? Click all such squares.



Item 14: F ALG-06-B

This rug has 6 rows and 6 columns. Each square will have a patch on it when it is finished, according to the answers to these questions:



What will this rug look like when it is finished? Click on the squares multiple times to change the patches.