



Co-funded by  
the European Union

## Development of the COMATH

The results are based on the work within the project “Computational Thinking and Mathematical Problem Solving, an Analytics Based Learning Environment” (CT&MathABLE). Coordination: Prof. Valentina Dagienė, Vilnius University (Lithuania). Partners: Ankara University (Türkiye), Eötvös Loránd University (Hungary), Gedminų Progymnasium (Lithuania), KTH Royal Institute of Technology (Sweden), Özkent Akbilek Middle School (Türkiye), University of Basque Country (Spain), University of Turku (Finland). The project has received co-funding by the Erasmus+ Programme KA220.

These results are developed by Daranee Lehtonen, Heidi Kaarto, Janica Kilpi, Kim Erola, and Marika Parviainen under WP3.

Reviewers: Vaida-Masiulionytė Dagienė, Vilnius University, Lithuania, and Arnold Pears, KTH Royal Institute of Technology, Sweden

CT&MathABLE project (2022-1-LT01-KA220-SCH-000088736) 2023 license granted.





## General overview

This report outlines the initial development of **COMATH**, an instrument designed to assess **computational thinking (CT)** and **algebraic thinking (AT)** skills in students aged **9–14**. The assessment tool has been structured for three distinct age groups: **COMATH1** for ages **9–10**, **COMATH2** for ages **11–12**, and **COMATH3** for ages **13–14**.

The primary objective of this development phase was to prepare the **COMATH** assessment instrument for **piloting** in six partner countries: **Finland, Hungary, Lithuania, Spain, Sweden, and Turkey**. By integrating a classification system based on national curricula with a **systematic literature review** of existing assessment instruments and empirical evaluation data, the instrument aims to provide educators with a robust framework for understanding and assessing these critical skills. Further testing and refinement will ensure its **effectiveness** in real-world educational settings.



## Target groups


This document provides an overview of the **development process** of the **COMATH** assessment instruments to support:

1. **Researchers** – to gain insights into the development of **high-validity** and **high-reliability** assessment instruments, which can be applied to future studies.
2. **Test developers** – to assist in designing tools for assessing students' **CT** and **AT** skills effectively.

### Keywords

algebraic thinking; computational thinking; assessment instrument; K–9 education; development

## Table of Contents

1. INTRODUCTION.....	4
2. AN OVERVIEW OF THE COMATH DEVELOPMENT.....	5
2.1 Systematic Literature Review and Learning Outcomes Identification .....	5
2.2 Classification System Development .....	5
2.3 Test Item Development .....	5
2.3 Expert Evaluation .....	5
2.4 Usability Testing .....	6
2.5 Finalising the Assessment Instrument .....	7
3. THE DEVELOPMENT OF CT TEST ITEMS.....	8
3.1 CT Classification for COMATH Test Items.....	8
3.2 Item Response Theory Analysis of Bebras Challenge 2022 Tasks .....	8
3.3 CT Expert Evaluation.....	9
3.4 Finalising CT Test Items .....	10
4. THE DEVELOPMENT OF AT TEST ITEMS .....	12
4.1 Classification of AT for COMATH Test Items .....	12
4.2 AT Test Item Development .....	16
4.3 AT Expert Evaluation .....	16
4.4 Finalising AT Test Items .....	17
5. COMATH 1–3 FOR THE FIRST PILOT STUDY.....	18
 REFERENCES .....	19

## 1. Introduction

The **COMATH** assessment instrument was developed to evaluate **computational thinking (CT)** and **algebraic thinking (AT high validity and reliability)** that could be used across different cultural contexts.

To account for students' varying stages of cognitive development, the assessment was designed for three distinct age groups:

- **COMATH1** for ages 9–10
- **COMATH2** for ages 11–12
- **COMATH3** for ages 13–14

The need for separate assessments arose from differences in students' learning progress. While younger students (COMATH1 and COMATH2) are in the early stages of developing CT and AT skills, older students (COMATH3) engage with more advanced concepts. Therefore, each COMATH instrument was designed to reflect these varying levels of complexity in digital skill development and accurately measure students' CT and AT abilities.

The development of COMATH took place between April and September 2023 at the Turku Research Institute for Learning Analytics, University of Turku, Finland, in collaboration with project partners from five other countries. To ensure its validity and reliability, the development process was:

- **Theoretically grounded** and guided by well-established design frameworks
- **Informed by experts** from multiple disciplines, including researchers from diverse cultural backgrounds and teachers working with different age groups
- **Based on both quantitative and qualitative research methods**

After its initial development, COMATH was tested in a pilot study conducted from October 2023 to February 2024 across six partner countries: Finland, Hungary, Lithuania, Spain, Sweden, and Turkey. The findings from this phase led to further refinements, paving the way for a comprehensive pilot test scheduled for autumn 2024, which will evaluate the instrument's validity and reliability.

## 2. An Overview of the COMATH Development

One of the project's aims was to develop high-reliability and validity assessment instruments for evaluating computational thinking (CT) and algebraic thinking (AT) skills among students aged 9–14. To meet that aim, the COMATH assessment instrument was developed based on (1) a classification system that linked task designs to specific learning outcomes and (2) feedback from international experts in the field. The development of COMATH was as follows:

### 2.1 Systematic Literature Review and Learning Outcomes Identification

At the beginning, we conducted a systematic literature review of existing CT and AT assessment instruments (see Report 3.1) to identify test items relevant to the assessment of CT and AT and what skills they aimed to measure. In parallel, we identified CT and AT learning outcomes that the national curriculum of the project partners' countries (Finland, Hungary, Lithuania, Spain, Sweden, and Turkey) aims to develop among 9–14-year-old students (see Report 2.1).

### 2.2 Classification System Development

We built the classification system for designing COMATH test items upon the defined learning outcomes and the systematic review. It linked specific task designs to corresponding sets of skills, aligning the key skills and competencies we intended to assess with the expected learning outcomes in the curricula. The classification system served as the foundation for the development of the assessment instruments.

### 2.3 Test Item Development

For the development of the test items, we utilised existing CT and AT assessment instruments that were found in the systematic review and demonstrated at least moderate psychometric quality. The CT test items were supplemented with Bebras 2022 tasks, which are widely recognised as reliable tools for CT skills. We also generated new test items when there were no existing test items that assessed mathematics with CT and AT skills. Due to the different development phases of the target students, the difficulty level of the assessment instrument needs to reflect the differences in the complexities of students' digital skill development and capture the level of their CT and AT skills. Therefore, the COMATH was tailored to three distinct age groups:

1. **COMATH1** for 9–10-year-olds
2. **COMATH2** for 11–12-year-olds
3. **COMATH3** for 13–14-year-olds

### 2.3 Expert Evaluation

After developing the test items, they were reviewed by experts in CT and AT from each of the partner countries. These experts included researchers, teacher educators, and qualified teachers with direct experience in teaching the relevant subjects. The purpose of this evaluation was to ensure the content validity of the test items—meaning that the items aligned with the curricula and educational contexts across the different countries involved in the project.

A total of 13 CT experts and 16 AT experts (at least two experts for each skill area per country) were asked to review the test items. They provided feedback on various aspects, such as:

- Whether the items were clear and easy to understand.
- Whether the items effectively assessed the intended skills.
- Whether the items were necessary and appropriate for the target age groups.
- Suggestions for potential improvements.

The experts' feedback, focusing on the clarity, relevance, and completeness of the items, was used to make further revisions. This ensured that the test items would be suitable for a variety of educational settings and contexts.

## 2.4 Usability Testing

The initial version of the assessment instrument was implemented in ViLLE, a digital learning platform. In parallel with the expert evaluation, the instrument was tested with seven Finnish students to evaluate its face validity, including its usability and understandability using a “thinking-aloud” method, in which the students verbalised their thoughts while doing the test in ViLLE.

The usability testing took place in September 2023, at a primary school in Southern Finland, using a selection of CT and AT test items from the COMATH. Participants included a 4th-grade girl, a 5th-grade boy, and an 8th-grade boy. The test session was planned to last one school lesson (45 minutes), with participants free to leave when they finished. Each participant did the COMATH test intended for their age. They were encouraged to raise their hands and ask questions if they encountered any difficulties.

The 4th grader completed the test in 28 minutes, followed by the 8th grader at 34 minutes, and the 5th grader at 40 minutes. Overall, the usability of the tool was smooth, with no major issues arising. However, the feedback gathered was somewhat limited. This may have been due to the fact that the participants took the test at the same time, did not know each other, and were unfamiliar with the organiser of the testing. As a result, they seemed hesitant to ask questions or offer feedback about the test items.

Despite the limited feedback, the insights we did receive were valuable. For instance, the youngest participant was initially unsure about how to submit her answers, suggesting that clearer instructions on this process are necessary. Additionally, some of the tasks, particularly the text-heavy sections, were found to be too challenging for the younger students, even though we had made efforts to minimize the amount of text. In terms of the overall setup and timing, everything went smoothly. The test duration and structure worked well, and the COMATH assessment instrument seemed to be fully ready for the project pilot study.

In addition to the first testing, the same subset of CT tasks was tested in another town in the area, with a 6th-grade boy. This student provided verbal feedback for each task. Combining his feedback with his digital trace data from the ViLLE learning analytics and results from the earlier tests gave us valuable insights into the student's perspective. This helped us finalise the task order for the upcoming pilot study.

Additionally, all the AT test items were also tested with three students in another town in Southern Finland. Each student represented one of the target age groups for COMATH1–3. The goal was to assess the clarity of the instructions and the time required to complete the test. Based on these test results, we made modifications to the AT test items, particularly in terms of instruction clarity and the number of items suitable for completion within a 40-45-minute lesson.

## 2.5 Finalising the Assessment Instrument

The final stage involved integrating the developed test items into a structured and cohesive set of assessment tasks to evaluate students' CT and AT skills. Each test was designed to be completed within 40–45 minutes, aligning with the typical duration of a single lesson in most partner countries. This time frame was carefully chosen to ensure the assessment could be administered efficiently within standard school schedules while maintaining its rigour and reliability in evaluating students' skills.

Following modifications based on received feedback, the assessment instruments were implemented in ViLLE, a digital learning environment, and subsequently translated from English into the eight official languages of the partner countries (Finnish, Hungarian, Lithuanian, Turkish, Swedish, Spanish, Basque, and Catalan). This ensured that the test would be accessible to students from diverse linguistic backgrounds and facilitated the first phase of the pilot study in each partner country with appropriate linguistic adaptations.

The test items were carefully designed to be age-appropriate and sufficiently challenging, incorporating CT and AT elements that aligned with the learning outcomes defined in the national curricula of the partner countries. Before the first pilot, the project team conducted a thorough review of the test items, correcting any typographical errors or technical issues to ensure accuracy and usability.

## 3. The Development of CT Test Items

### 3.1 CT Classification for COMATH Test Items

Following our systematic literature review on the definitions and classifications of CT (See Report 3.1), we concluded that CT is a broad and complex concept comprising multiple interconnected sub-concepts. Among the most recent literature (Ezeamuzie & Leung, 2022; Shin et al., 2022), we selected Shute et al.'s (2017) definition of CT because (1) it is derived from a comprehensive review of CT in education, and (2) it is well-suited to our purpose of annotating tasks for the development of COMATH.

Shute et al. (2017, p.151) define CT as “the conceptual foundation required to solve problems effectively and efficiently (i.e., algorithmically, with or without the assistance of computers), with solutions that are reusable in different contexts.” According to Shute et al. (2017, pp.153), CT consists of six key components:

1. **Decomposition** - Breaking down a complex problem or system into smaller, manageable parts. These parts are functional elements that work together to form the whole.
2. **Abstraction** – Identifying the core aspects of a system. This includes:
  - **Data Collection & Analysis** – Gathering relevant data from various sources and understanding their relationships.
  - **Pattern Recognition** – Detecting patterns or underlying rules within the data.
  - **Modelling** – Creating simulations or models to represent a system’s behavior or predict future outcomes.
3. **Algorithms** – Developing structured and logical steps to solve problems, which can be executed by humans or computers. This involves:
  - **Algorithm Design** – Creating step-by-step solutions.
  - **Parallelism** – Performing multiple steps simultaneously.
  - **Efficiency** – Optimizing the process by eliminating unnecessary steps.
  - **Automation** – Enabling solutions to run automatically for repeated tasks.
4. **Debugging** – Identifying and fixing errors when a solution does not function correctly.
5. **Iteration** – Refining solutions through repeated testing and improvement until the desired outcome is achieved.
6. **Generalisation** – Applying CT skills across different situations and domains to solve various problems efficiently.

### 3.2 Item Response Theory Analysis of Bebras Challenge 2022 Tasks

We conducted an *Item Response Theory* (IRT) analysis of the tasks used in the 2022 Bebras Challenge (<http://bebras.org>). The Bebras tasks are designed to engage students with computer



science and CT, spark their curiosity, and promote a deeper understanding of technology (Araujo et al., 2019; Dagienė & Sentence, 2016). Our aim was to identify tasks from the Bebras Challenge that could be incorporated into the COMATH assessment, particularly those that effectively differentiate students based on their CT skills. Data from 88,041 students in Lithuania and Hungary were analysed for this purpose. By examining item difficulty and discrimination in relation to different age groups, we assessed the tasks at the item level to ensure the instruments functioned as intended. Additionally, percentile norms were calculated for each country to provide a basis for interpreting individual student results.

In IRT, *discrimination* (a) refers to how effectively a task distinguishes between students with different skill levels. A highly discriminative task accurately differentiates between students with varying levels of proficiency, making it more effective for assessing a broad range of abilities within a given age group. Tasks with higher discrimination values are preferred, as they provide deeper insights into students' skill levels.

A steeper ICC curve indicates better discrimination between students, while a curve positioned further to the right represents a more challenging task. Additionally, curves that start higher on the y-axis take into account the probability of guessing correctly. For example, the tasks with the steepest curves—indicating the highest discrimination—were 2022-CH-08 and 2022-BR-01.

Ultimately, we selected 41 Bebras tasks with strong discriminatory power for inclusion in the COMATH CT assessment. We then analysed the difficulty level of each selected task.

*Item difficulty* (b) refers to the point where the ICC has the steepest slope. The higher the difficulty of a task, the greater the level of ability required for a student to answer it correctly. Tasks with high b values (greater than 1) are considered very difficult, meaning that students with lower ability levels are unlikely to answer them correctly. Conversely, tasks with low b values (below -1) are classified as easy, allowing most students, including those with lower ability, a reasonable chance of answering them correctly. Tasks with b values between -0.5 and 0.5 are regarded as having a medium difficulty level.

### 3.3 CT Expert Evaluation

In addition to the IRT analysis, we evaluated the *content validity* of the selected Bebras tasks to ensure their relevance and representativeness in assessing CT skills for the development of COMATH test items. Content validity refers to the extent to which an assessment instrument effectively measures the targeted construct for a specific purpose (Almanasreh et al., 2019).

To achieve this, we consulted 13 CT experts from the project partners' countries, each with an average of 17 years of experience in CT education. The experts were asked to rate how well each selected task measured specific CT skills based on Shute et al.'s (2017) classification. They assigned ratings on a scale of “Well”, “Somewhat”, or “Not at all”. We then calculated the *content validity ratio* (CVR) using the following formula:

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

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, first proposed by Lawshe (1975), is among the most commonly used statistical techniques for quantifying content validity. The CVR ranges from -1 to 1, with higher values indicating greater consensus among panel members regarding the necessity of an item in an instrument. The CVR value is determined using the Lawshe Table (Lawshe, 1975). For instance, in this study, where the panel consists of 13 members, an item is considered acceptable at a significant level if its CVR exceeds 0.54.

The CVR analysis revealed that most tasks primarily focused on assessing algorithmic thinking, although some also evaluated abstraction skills. These findings align with Araujo et al. (2019), who examined the Bebras Challenge and observed that Bebras tasks often integrate multiple CT skills, with algorithmic thinking being a core component. Based on these insights, we categorised 41 Bebras tasks selected for COMATH into two groups:

1. Tasks that assess algorithmic thinking exclusively.
2. Tasks that require algorithmic thinking alongside additional CT skills.

Furthermore, experts provided qualitative feedback on various aspects of the tasks, including content, structure, suitability for different age groups (9–10, 11–12, and 13–14 years), and potential cultural sensitivities. These expert insights were subsequently used to refine the test items.

### 3.4 Finalising CT Test Items

#### Task Difficulty and Visualisation

After dividing the selected tasks into two groups—algorithmic thinking and algorithmic thinking combined with additional CT skills—we positioned the tasks within each group on the same number line (a difficulty axis) according to their difficulty level for each age group, based on the IRT analysis. The difficulty scale ranged from -2 to 2 for each COMATH age group, ensuring that tasks aligned with the skill levels of the targeted students.

#### Task Selection and Difficulty Distribution

Once the tasks were grouped by similar difficulty levels, we selected the task with the best discrimination capability from each group for inclusion in COMATH. However, as shown in Figures 10–15, the selected tasks were not evenly distributed across the full range of skill levels within each age group. Since a well-constructed assessment instrument should contain test items of varying difficulty, we designed additional tasks to fill the gaps and ensure a balanced distribution of easy, medium, and difficult tasks along the difficulty axis.

To create these new tasks, we estimated their difficulty levels based on the 2022 Bebras Challenge task analysis, ensuring they complemented the existing tasks. This process included:

- **Modifying existing Bebras tasks** (e.g., shortening text or reducing the number of multiple-choice answers to decrease difficulty).
- **Adjusting tasks across age groups** (e.g., using a medium-difficulty task from an older age group as a difficult task for a younger group and vice versa).

- **Designing entirely new tasks** inspired by test items identified in our systematic literature review (see Report 3.1).

### Final Task Set and Variations

Table 1 provides all CT test items included in COMATH 1–3 for the first pilot study. In this final version, we compiled a total of **29 CT test items**, including:

- 14 tasks assessing only algorithmic thinking.
- 15 tasks assessing algorithmic thinking alongside other CT skills.
- 23 tasks with an alternative 'B' version, where A and B versions were content-wise similar but featured minor variations (e.g., different images, rotated or reorganised visuals, or slight text modifications). These variations may influence Pilot 1 results, and their analysis will be used to refine the COMATH assessment instrument.
- 18 anchor tasks, with 8 included across all levels of COMATH (COMATH 1–3).

## 4. The Development of AT Test Items

### 4.1 Classification of AT for COMATH Test Items

#### Definition of AT

The classification of Algebraic Thinking (AT) test items in COMATH was based on the systematic review of existing AT assessment instruments (see Report 3.1). This review provided a foundation for developing AT test items, ensuring alignment with established frameworks and definitions of AT.

Our review highlighted a shift in focus from specific algebraic content to students' algebraic thinking processes, which are essential for building competency in understanding and applying algebra. Algebraic Thinking (AT) is a cognitive process that involves making sense of algebraic concepts. It encompasses (e.g., Kaput, 2008):

1. Identifying and generalising mathematical structures and relationships
2. Representing generalisations using alphanumeric symbols and other representations, such as diagrams and graphs
3. Reasoning and modelling with symbolised generalisations

The findings suggest that AT should be viewed as a multi-component construct rather than a single skill.

#### Developing an AT Classification for COMATH

To determine the AT skills to be assessed in COMATH, we examined the contents of existing assessment instruments and identified recurring AT competency areas from the reviewed studies. We also referenced AT classifications from the academic literature cited in these studies. Through this process, we established seven key AT competency areas:

1. Generalised Arithmetic
2. Equations and Inequalities
3. Functional Thinking
4. Variables
5. Representation
6. Transformation
7. Transversal Skills

However, since the ability to work with variables (i.e., symbols, typically letters representing generalised or unknown values in mathematical relationships, treated as numbers) is fundamental to all other AT competencies, we chose not to classify it as a standalone competence. Instead, it is incorporated within the other AT skills assessed in COMATH.

## Final AT Classification for COMATH

As a result, the COMATH AT test items are categorised into six key competency areas, detailed below:

### 1. Generalised Arithmetic

The ability to identify and extend arithmetic relationships, including fundamental properties of numbers and operations (e.g., the Commutative Property of Addition). This also involves reasoning about the structure of arithmetic expressions rather than focusing solely on their computational outcomes (Blanton et al., 2015).

#### 1.1 Efficient Numerical Manipulation (Procedural Flexibility)

The ability to simplify calculations by leveraging number relationships and compensation strategies, allowing for more efficient problem-solving without relying on direct computation.

#### 1.2 Generalisation

The ability to recognise and apply common mathematical properties (e.g., odd/even numbers, doubling, commutativity) and identify patterns in arithmetic operations. This includes understanding that a mathematical procedure effective in one equation may also be applicable to another.

### 2. Equivalence, Equations, and Inequalities

The ability to understand the **equal sign** relationally, recognising that it represents equivalence between two math expressions rather than simply indicating an answer. This competency also involves reasoning with symbolic expressions and equations, as well as describing relationships between generalised quantities, regardless of whether they are equivalent or not (Blanton et al., 2015).

#### 2.1 Understanding the Equal Sign (*Name and Definition*)

Students should be able to name the equal sign and correctly define and explain its relational meaning—that it signifies that both sides of an equation hold the same value. This contrasts with an operational interpretation, where the equal sign is mistakenly viewed as a signal for the result of an arithmetic operation (addition, subtraction, multiplication, or division) or merely placed before an answer.

#### 2.2 Open Number Sentences

The ability to determine an unknown whole number in addition, subtraction, multiplication, and division equations—particularly in early education. Developing reasoning skills with open number sentences supports broader AT competencies, such as understanding the properties of zero, making conjectures, and constructing mathematical justifications.

#### 2.3 Working with Pictorial Variables (*Applicable only to COMATH 1–2*)

The ability to interpret and manipulate visual representations of variables, such as shapes, images, or icons, used as symbolic placeholders for unknown values in equations and inequalities. This includes using algebraic reasoning to simplify, substitute, and solve for unknown values.

## **2.4 Working with Letter Variables**

The ability to interpret, manipulate, and apply letter symbols as representations of unknown or generalised values in equations and inequalities. This includes understanding how variables function within mathematical expressions and using algebraic reasoning to simplify, substitute, and solve for unknown values.

## **2.5 Solving Word Problems**

The ability to translate a word problem into a mathematical statement, such as an equation or inequality, and then solve it. This involves identifying relevant information, formulating an appropriate algebraic representation, and applying problem-solving strategies.

# **3. Functional Thinking**

Functional thinking involves the ability to identify, generalise, and describe numerical and figural patterns to understand relationships between co-varying quantities. This includes recognising similarities, differences, causality, and patterns of growth (Kaput, 1998). It requires students to observe patterns, establish rules governing those patterns, and apply them to extend, predict, or generate new sequences. Patterns can be classified as linear or nonlinear, depending on how they progress.

## **3.1 Figural Patterns**

The ability to recognise, describe, extend, and create patterns involving shapes, figures, or diagrams. This includes identifying repeating sequences, transformations, and growth patterns within visual representations.

## **3.2 Numerical Patterns**

The ability to analyse and extend numerical sequences by recognising rules governing their progression. This involves understanding both arithmetic (linear) and geometric (nonlinear) patterns, making generalisations, and predicting future terms in a sequence.

## **3.3 Function Machines and Rules**

The ability to interpret, relate, and generate representations of functional relationships in various forms, such as tables, graphs, equations, and verbal descriptions. This includes identifying properties of linear functions (e.g., slope and intercepts), recognising simple nonlinear functions, and applying functions to solve real-world problems. Students should be able to translate between different representations of functions and use function machines to model input-output relationships effectively.

# **4. Representation**

Representation refers to the ability to interpret, construct, and utilise multiple forms of representation, including diagrams, tables, symbols, and verbal descriptions, to organise

information and develop a deeper understanding of mathematical relationships (Dindyal, 2003). Mastering different forms of representation helps students visualise abstract concepts, identify patterns, and establish connections between various mathematical ideas.

#### **4.1 Diagrammatic Representations**

The ability to interpret and construct visual representations of mathematical relationships, such as graphs, charts, geometric diagrams, and schematic illustrations. Diagrammatic representations help in modelling relationships, identifying patterns, and simplifying complex information. This skill is essential for problem-solving, reasoning, and effectively communicating mathematical ideas.

#### **4.2 Verbal and Symbolic Representations**

The ability to communicate mathematical relationships and problem-solving processes through written or spoken language and mathematical symbols. This encompasses explaining patterns, describing functions, reasoning with algebraic expressions, and justifying mathematical arguments. Proficiency in verbal and symbolic representation enables students to transition between words and symbols, articulate their reasoning clearly, and participate effectively in mathematical discussions.

### **5. Transformation**

Transformation refers to the ability to manipulate and restructure algebraic expressions, such as equations, while maintaining their equivalence. This involves applying various operations, such as combining like terms, factoring, expanding, substituting, performing polynomial operations, exponentiation, and simplification of mathematical expressions (Kieran, 1996). Mastery of transformation skills enables students to simplify, compare, and evaluate algebraic expressions efficiently, which is essential for solving equations, recognising patterns, and understanding mathematical structures and relationships.

#### **5.1 Equivalent Expressions**

The ability to rewrite algebraic expressions in an equivalent form to simplify calculations or facilitate problem-solving, such as solving equations. This includes expanding expressions, factoring, combining like terms, and using mathematical properties (e.g., distributive property) to transform mathematical expressions into more manageable forms. Recognising and generating equivalent expressions allows students to solve equations efficiently while ensuring the underlying relationships remain unchanged.

#### **5.2 Transforming Letter Variables**

The ability to manipulate algebraic expressions involving letter variables while preserving their mathematical relationships. This includes substituting values for variables, rearranging formulas, and rewriting expressions to highlight different aspects of an equation. Developing this skill enables students to work flexibly with symbolic representations, transition between different equation forms, and apply algebraic reasoning to solve problems effectively.



## 6. Transversal Skills for AT

Transversal skills encompass a broad set of higher-order cognitive abilities essential for applying AT effectively. These skills involve reasoning, generalisation, justification, modelling, prediction, validation, and problem-solving. These skills support flexible thinking and deep mathematical understanding, allowing students to apply AT effectively in different contexts.

- **Reasoning & Generalisation** – Identifying patterns, making logical deductions about mathematical relationships, and extending them to other contexts.
- **Justification & Proof** – Constructing and validating mathematical arguments.
- **Modelling & Prediction** – Representing real-world situations mathematically (e.g., equations, graphs, or functions) to analyse and solve problems as well as anticipating outcomes.
- **Validation** – Ensuring accuracy and reliability of mathematical solutions and reasoning processes
- **Problem-Solving** – Applying AT to identify, analyse, and solve complex problems.

### 4.2 AT Test Item Development

The development of Algebraic Thinking (AT) test items in COMATH was based on a systematic review of existing assessment instruments and test items. This process incorporated insights from IRT analysis, focusing on the difficulty and discrimination of test items used in the reviewed studies.

To ensure the validity and effectiveness of the assessment, we selected test items with strong discrimination ability and appropriate difficulty levels as the foundation for developing new items. These items were carefully adapted and refined to align with the Final AT Classification for COMATH (see Section 4.1). By building upon evidence-based practices, we ensured that the developed test items accurately measured key AT skills while maintaining consistency with AT literature and the national curricula of the project partner countries.

### 4.3 AT Expert Evaluation

Following the development of the AT test items, a rigorous expert evaluation process was conducted to ensure their content validity, clarity, and appropriateness for the target age groups and educational contexts of the partner countries. A panel of 16 AT experts, each with an average of 20 years of experience, was assembled from the project's partner countries to review the test items comprehensively.

The expert evaluation focused on several key aspects:

- **Relevance and Clarity** – Assessing whether each test item effectively measured the intended AT skills and whether the wording and structure were clear and comprehensible.
- **Alignment with Defined AT Skills** – Ensuring that the test items accurately reflected the AT classification framework and measured the skills as intended.



- **Content and Language Appropriateness** – Evaluating whether the language, phrasing, and level of difficulty were suitable for students aged 9–10, 11–12, and 13–14 years as well as educational contexts of the partner countries.
- **Mathematical Symbols and Notation** – Verifying that the symbols and mathematical conventions used in the test items were consistent with those taught in primary and lower secondary education across all partner countries.

This expert review was critical in refining the assessment items, ensuring that they were valid, reliable, and culturally appropriate for use across different educational systems.

#### 4.4 Finalising AT Test Items

Following expert feedback, the AT test items were carefully modified and refined to improve their accuracy, clarity, and effectiveness in assessing the targeted AT skills. The revisions aimed to ensure that each item was appropriately challenging, well-structured, and aligned with the AT skills being measured, as well as the educational contexts in which the first pilot study would be conducted.

Figures 16–31 present examples of AT test items included in COMATH 1–3 for the first pilot. In the finalised version, a total of 65–77 AT test items were compiled, covering a range of core AT skills:

- **Generalised Arithmetic:** 23, 25, and 25 items in COMATH 1–3, respectively.
- **Equivalence, Equations, and Inequalities:** 19, 21, and 22 items in COMATH 1–3, respectively.
- **Functional Thinking:** 9, 10, and 11 items in COMATH 1–3, respectively.
- **Representation:** 3, 4, and 5 items in COMATH 1–3, respectively.
- **Transformation:** 5, 8, and 10 items in COMATH 1–3, respectively.
- **Transversal Skills:** 7, 8, and 6 items in COMATH 1–3, respectively.

To evaluate different test item options for each AT skill and select the most effective items for the final assessment instrument, almost all test items included an alternative 'B' version. This version was identical to the 'A' version, except for the numerical values used. These variations enable an analysis of how changes in numbers affect item difficulty and discrimination, contributing to the refinement of the COMATH assessment instrument.

Additionally, a series of anchor tasks was included across two or all three levels of COMATH to maintain continuity and enable the comparison of students' skills across different age groups.

## 5. COMATH 1–3 for the First Pilot Study

Following its initial development, we conducted a pilot study to evaluate the effectiveness of the COMATH assessment instrument. This study involved over 3,000 students and their teachers across six partner countries—Finland, Hungary, Lithuania, Turkey, Sweden, and Spain—between autumn 2023 and spring 2024.

The first pilot study aimed to assess how well the test items functioned in measuring students' CT and AT skills across different age groups and educational contexts. To achieve this, we will analyse the results for each age group within each country, for all age groups within each country, and across all participating countries. This multi-level analysis will help determine the extent to which each test item effectively differentiates between varying skill levels. It will also provide insights into the discriminatory power of each task, identifying which items are most effective in assessing specific skill levels.

The findings from the first pilot study will guide further refinements and necessary modifications to COMATH, ensuring that the assessment instrument is optimally designed before the full pilot phase in autumn 2024. The full pilot will serve to validate the instrument's reliability and effectiveness across diverse educational settings, ultimately confirming its suitability for assessing students' CT and AT skills in the target age groups.



## 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>
- 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>
- Blanton, M., Stephens, A., Knuth, E., Gardiner, A. M., Isler, I., & Kim, J.-S. (2015). The development of children's algebraic thinking: The impact of a comprehensive early algebra intervention in third grade. *Journal for Research in Mathematics Education*, 46(1), 39–87. <https://doi.org/10.5951/jresmetheduc.46.1.0039>
- Dagienė V., & Sentence, S. (2016). “It is computational thinking! Bebras tasks in the curriculum,” *Proceedings of the International Conference on Informatics in Schools: Situation, Evolution, and Perspectives*, 28–39.
- Dindyal, J. (2003). *Algebraic thinking in geometry at high school level: Students' use of variables and unknowns* [Unpublished doctoral dissertation]. Illinois State University.
- 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.
- Kaput, J. (1998). Transforming algebra from an engine of inequity to an engine of mathematical power by “algebrafying” the K–12 curriculum. In National Council of Teachers of Mathematics, Mathematical Sciences Education Board, & National Research Council (Ed.), *The nature and role of algebra in the K–14 curriculum: Proceedings of a National Symposium* (pp. 25–26). National Academies Press.
- Kaput, J., Blanton, M., & Moreno, L. (2008). Algebra from a symbolization point of view. In J. Kaput, D. Carraher, & M. Blanton (Eds.), *Algebra in the early grades* (pp. 19–55). Lawrence Erlbaum Associates.
- Kieran, C. (1996). The changing face of school algebra. In C. Alsina, J. Alvarez, B. Hodgson, C. Laborde, & A. Pérez (Eds.), *The 8th International Congress on Mathematical Education: Selected lectures* (pp. 271–290). Thales.
- 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>
- Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review*, 22, 142–158.