Co-funded by
the European Union

Further development of COMATH based on feedback

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 and Marika Parviainen under WP3.

Revised by Valentina Dagienė

General overview

This report outlines the further development of COMATH, an instrument designed to assess computational thinking (CT) and algebraic thinking (AT) skills in students aged 9–14. The improvements build upon the results of the first pilot (see Report R3.3), which aimed to collect feedback from teachers and students in six partner countries (Finland, Hungary, Lithuania, Turkey, Sweden, and Spain). The primary objective of the pilot was to evaluate how well the test items functioned in measuring students' CT and AT skills across different age groups and educational contexts.

To guide the refinement process, multiple data sources from the first pilot were analysed, including students' test results and feedback from both students and teachers collected through surveys and interviews. The insights gained from these sources informed necessary modifications and enhancements to COMATH. These refinements aimed to optimise the assessment, ensuring that both the CT and AT tests remain valid and reliable while being practical for classroom implementation—allowing completion within a single class lesson. The revised version underwent the second pilot in autumn 2024.

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

One of the project's key aims was to develop a high-reliability and high-validity assessment instrument for evaluating computational thinking (CT) and algebraic thinking (AT) skills among students aged 9–14. The assessment was structured into three distinct test versions, each tailored to a specific age group: Age Group 1 (9–10), Age Group 2 (11–12), and Age Group 3 (13–14). To achieve this, the COMATH assessment instrument underwent further development in spring 2024. The refinement process was guided by multiple data sources from the first pilot, including students' test results and feedback from both students and teachers collected through surveys and interviews (see Report R3.3).

The insights gained from these sources informed essential modifications and enhancements to COMATH. The time students spent on completing the CT and AT tests during the first pilot was analysed to estimate and determine the appropriate number of test items for the second pilot. Based on Item Response Theory (IRT) analysis, test items with appropriate discrimination power and difficulty levels—ensuring an even distribution from easy to difficult—were retained. Redundant items were removed, and the content of some items was adjusted either to improve their discrimination and difficulty balance or to reduce the time required for completion. These refinements aimed to optimise the assessment, ensuring both the CT and AT tests remained valid, reliable, and practical for classroom implementation—allowing completion within a single 40-minute lesson.

As a result of these revisions, the CT and AT tests were consolidated into a single version, replacing the two versions (A and B) used in the first pilot. The number of CT items per age group was reduced from 18–19 (The first pilot) to 14, while AT items were streamlined from 66–79 per age group (The first pilot) to 54–63. The revised version will undergo the second pilot in autumn 2024.

2. The Development of Computational Thinking (CT) Test

2.1 Completion Time

In the first pilot, students completed the CT and AT tests in separate lessons. Each test was originally designed to be completed within 45 minutes. However, students were given unlimited time to finish the tests. The time students spent on completing the CT and AT tests during the first pilot was analysed to estimate and determine the appropriate number of test items for the second pilot, ensuring that each test could be completed within a single 40-minute lesson.

Overview

Table 1 presents the average time students in all age groups spent completing the CT test, which consisted of 18 items for Age Groups 1 and 2, and 19 items for Age Group 3. The mean completion times were 24 minutes for Age Group 1, 23 minutes for Age Group 2, and 26 minutes for Age Group 3, with standard deviations of 10, 8, and 8 minutes, respectively. While students in Age Groups 1 and 2 spent a similar amount of time on the test, students in Age Group 3 took slightly longer. The longest completion time recorded was 61 minutes in Age Group 1, followed by 59 minutes in Age Group 3 and 53 minutes in Age Group 2. Figure 1 (available in a full version of the report) illustrates the distribution of students based on time spent on the CT test. Among the 3,350 students who participated in the first pilot, 55% (n = 1,840) completed the test within 18–30 minutes, while only 3.64% (n = 122) spent more than 40 minutes on the test.

Table 1. Mean, standard deviation (SD), and maximum completion time (in minutes) for the CT test across age groups

(min)	Age Group 1	Age Group 2	Age Group 3	Age Groups 1–3
Mean completion time	23.51	23.35	26.01	24.13
SD of Completion time	9.80	7.95	8.29	8.75
Maximum completion time	61.37	53.33	58.88	61.37

Distribution of Time Spent on Each Section

In the first pilot, the CT test comprised two sections: Section 1, which assessed algorithmic thinking skills, and Section 2, which evaluated algorithmic thinking alongside other CT skills. The test versions for Age Groups 1 and 2 contained 18 items, with 9 in each section. The version for Age Group 3 included 19 items, with 9 in Section 1 and 10 in Section 2 (see Report 3.2 for details).

As shown in Table 2, students spent approximately 12 minutes on Section 1 and 13 minutes on Section 2. When examining time distribution by age group, students in Age Groups 1 and 2 spent nearly equal time on both sections, averaging 11–12 minutes per section. However, students in

Age Group 3 allocated less time to Section 1 (11 minutes) than to Section 2 (15 minutes), which is reasonable given that Section 1 contained one item fewer than Section 2.

Table 2. Mean completion time (in minutes) for each section of the CT test across age groups

Mean completion time (min)	Age Group 1	Age Group 2	Age Group 3	Age Groups 1–3
Algorithmic thinking skills	11.932	11.399	11.348	11.559
Algorithmic thinking and other CT skills	11.580	11.947	14.664	12.575

Figures 2 and 3 (available in a full version of the report) illustrate the distribution of students based on the time spent on each section. The majority of students completed Section 1 within 8–14 minutes (53%) and Section 2 within 9–16 minutes (52%), indicating a relatively consistent time allocation across sections. To ensure that all students can complete the assessment comfortably within a standard 40-minute lesson for future classroom implementation, we proposed a maximum 20-minute time limit for each section. A smaller proportion of students took significantly longer, with only 4% (n = 139) spending more than 20 minutes on Section 1 and 7% (n = 236) exceeding 20 minutes on Section 2. This suggests that while most students managed to complete both sections within a similar time range, a subset required additional time, particularly for Section 2, which contained more items in the Age Group 3 version.

We conducted a detailed analysis of the time distribution for students (N = 3,350) across both sections of the CT test. This analysis focuses on key statistical measures, including the median, interquartile range (IQR, 25th–75th percentile), and the 10th–90th percentile range, to better understand how students allocated their time. The results for each section are as follows:

Section 1: Algorithmic thinking skills

The **median** time spent on Section 1, representing the middle value in the data set when arranged in ascending order, was approximately 11–12 minutes. This means that half of the students completed the section in less time, while the other half took longer. Notably, the median was equal to the mean, indicating a relatively symmetric distribution of time spent.

The **IQR** capturing the middle 50% of students, ranged from approximately 8–9 minutes (25th percentile) to 14–15 minutes (75th percentile). This suggests that the majority of students completed Section 1 within this time frame, reflecting moderate variability in time allocation.

The **10th to 90th percentile range**, covering the middle 80% of the data, spanned approximately 6 to 18 minutes. This indicates that most students required between 6 and 18 minutes to complete the section, with only a small number taking significantly less or more time.

Section 2: Algorithmic thinking, along with other CT skills

The **median** time spent on Section 2 was slightly higher than for Section 1, at around 12–13 minutes. This means that half of the students completed the section in less time, while the other half took longer. The mean, however, differed slightly from the median, falling within the range of 12–15 minutes, suggesting a slightly skewed distribution.

The **IQR** extended from approximately 9–10 minutes (25th percentile) to 15–16 minutes (75th percentile). This indicates that most students took between 9 and 16 minutes to complete Section 2. Compared to Section 1, the distribution was slightly shifted toward longer completion times, suggesting that students generally required slightly more time for this section.

The **10th to 90th percentile range** for Section 2 spanned approximately 6 to 20 minutes, meaning that most students required between 6 and 20 minutes to complete the section. The slightly wider spread compared to Section 1 suggests greater variability in time allocation.

Summary

Overall, the distribution of time spent on Sections 1 and 2 reveals consistent patterns in student time allocation. The similarity between the mean and median in both sections indicates a balanced central tendency in time spent. The IQR analysis shows that the middle 50% of students spent between 8–15 minutes on Section 1 and 9–16 minutes on Section 2, indicating moderate variability. The 10th–90th percentile range suggests that most students completed Section 1 within 6–18 minutes and Section 2 within 6–20 minutes. The slightly wider range in Section 2 indicates greater variability in time spent. This may be because the section required students to apply a broader set of CT skills in addition to algorithmic thinking. Furthermore, in the version for Age Group 3, Section 2 contained one additional item compared to Section 1. Despite this variability, the overall time distribution remains concentrated around the middle values, with only a small proportion of students taking significantly less or more time.

Recommendations for Test Revision

In the first pilot, the CT test consisted of 18–19 items, which resulted in variability in completion times among students. This variation poses challenges for future classroom implementation, as some students may need more time to complete the test than others. To address this issue and enhance feasibility, we decided to slightly reduce the number of test items for age group. The goal was to ensure that all students can complete the assessment comfortably within a standard 40-minute lesson for future classroom implementation, making it more practical for integration into regular classroom activities. The final number of items will be determined based on detailed statistical analyses, specifically, Confirmatory Factor Analysis (CFA) and Item Response Theory (IRT), to ensure the test remains reliable and effective.

2.2 Quality Analysis of Test Items

Overview

Statistical analyses were carried out to assess the quality of the CT test items used in the first pilot. For detailed statistical results, please refer to Report 3.3.

The evaluation process began with a **Confirmatory Factor Analysis (CFA)**. This analysis explored the relationships (correlations) among test items to determine whether they measured the same CT skills within each age group. Items that showed low correlations with others were considered misaligned and were removed from the test.

Following the CFA, a **two-parameter Item Response Theory (IRT) analysis** was conducted. This method provided insights into two key properties of each test item: *item discrimination* and *item difficulty*. The results guided the selection and revision of items, with the aim of developing a reliable and valid assessment instrument. Such a tool can effectively map students along a continuum of the targeted latent trait—in this case, CT skills. By including items that vary in difficulty and are capable of distinguishing between students with different skill levels, the test becomes more precise. At the same time, this approach reduces the number of items needed, helping to avoid unnecessary burden on students.

Item Discrimination

Item discrimination measures how well a test item can differentiate between students with high, medium, and low levels of ability. Items with **positive discrimination values** indicate that higher-ability students are more likely to answer correctly, which aligns with expected patterns—namely, that the probability of a correct response increases as a student's ability level increases. Items demonstrating this pattern were retained in the COMATH test for the second pilot.

Only items with moderate to high discrimination values (at least 0.65) were considered suitable. Items with low discrimination values (below 0.65) lacked the ability to distinguish students effectively and were therefore excluded.

Items with **negative discrimination values** were also dropped. These suggest an unexpected pattern: students with lower ability levels were more likely to answer correctly than those with higher ability. Such results may indicate problems with item clarity or misleading information that disproportionately affected higher-performing students.

In rare cases, items with **very high discrimination values** (above 4) were examined more closely. Although such values may point to model instability, they do not automatically indicate poor item quality. Final decisions about whether to keep or revise these items were made based on both their statistical properties and expert review of the content.

Item Difficulty

Item difficulty refers to the level of ability a student needs to respond to a particular item correctly. The higher the difficulty value, the greater the level of ability required. A well-designed test should include items spanning a wide range of difficulty levels to ensure it is suitable for all students within the target group.

According to established guidelines (e.g., Baker, 2001; Gyamfi & Acquaye, 2023), acceptable difficulty values generally range from -3 to 3. Items with values outside this range were considered overly easy or overly difficult and were removed from the CT test, as they could reduce the test's overall effectiveness.

For each age group, the remaining items were positioned along the ability scale based on their difficulty values. The aim was to include a balanced number of items across three difficulty levels within the range of -2 to 2:

- **Easy**: Difficulty values between -2 and -1
- **Moderate**: Difficulty values between -1 and 1
- **Difficult**: Difficulty values between 1 and 2

When two versions of an item (A and B) were available, the version that contributed more effectively to the even distribution of difficulty levels was retained. The other was excluded.

In cases where the distribution of item difficulty was uneven, some items were revised to improve balance. This helped ensure that the test provided a comprehensive assessment across all ability levels.

Results and Actions Taken

The total number of items was reduced from 18–19 to 14 items for all age group. Each section now consists of seven items, maintaining a balanced assessment of CT skills while improving practicality for classroom use.

Age Group 1 (Students Aged 9–10)

The CFA revealed that five items—ALG-07-A, ALG-07-B, OTH-02, OTH-07-A, and OTH-07-B—showed weak correlations with other items. Additionally, ALG-07-A, OTH-07-A, and OTH-07-B had very low discrimination values, while ALG-07-B had an overly high value. These items were removed from the CT test. Two more items (ALG-10-A and ALG-10-B) were excluded due to their high difficulty levels.

The remaining 26 items were positioned along the ability scale based on their difficulty values. Item difficulties of both test sections were skewed towards the higher end, indicating that the test was particularly demanding for lower-performing students.

To improve the balance in item number, content, and difficulty, items for the second pilot were selected based on their discrimination values, difficulty levels, and content relevance. Three items were revised to adjust their difficulty: ALG-13-B and OTH-03-B were made easier, while OTH-08 was made more challenging.

Age Group 2 (Students Aged 11–12)

Low correlations with other items led to the removal of six items: ALG-07-A, ALG-07-B, ALG-08-A, ALG-08-B, OTH-07-A, and OTH-07-B. In addition, these items showed either poor discrimination (ALG-07-A) or difficulty levels that were too low (ALG-08-A and ALG-08-B) or too high (ALG-07-A and ALG-07-B).

The remaining 27 items were mapped onto the ability scale based on their difficulty values. Figure 6 shows (available in a full version of the report) that in Section 1, which assesses algorithmic thinking skills, item difficulties were unevenly distributed: about half the items were moderately easy (ranging from -0.5 to 0), while the other half were more difficult (1 and above). Figure 7 (available in a full version of the report) indicates that Section 2, assessing algorithmic thinking alongside other CT skills, contained items that were generally more difficult, especially for lower-performing students.

As with Age Group 1, items for the second pilot were selected based on their discrimination, difficulty, and content relevance. Four items were revised to adjust their difficulty: ALG-03-A, ALG-13-B, and OTH-13-B were made easier, while OTH-08 was made more challenging.

Age Group 3 (Students Aged 13–14)

Four items were removed from the CT test based on different criteria: low correlations with other items (ALG-05-A, ALG-05-B, and OTH-05-A), low discrimination (OTH-09-B), overly high discrimination (ALG-05-A), and very high difficulty levels (ALG-05-A and ALG-05-B).

The remaining 30 items were mapped along the ability scale. The distribution of item difficulties was more even within the range of -2 to 2 compared to the younger age groups. This suggests that the difficulty levels were better matched to the performance levels of students in this group.

Items for the second pilot were selected using the same criteria: discrimination, difficulty, and content relevance. No items were revised.

3. The Development of Algebraic Thinking (AT) Test

Based on Item Response Theory (IRT) analysis, test items with appropriate discrimination power and difficulty levels—ensuring an even distribution from easy to difficult—were retained. Redundant items were removed, and the content of some items was adjusted either to improve their discrimination and difficulty balance or to reduce the time required for completion.

3.1 Completion Time

In the first pilot, students completed the CT and AT tests in separate lessons. Each test was originally designed to be completed within 45 minutes. However, students were given unlimited time to finish the tests. The time students spent on completing the CT and AT tests during the first pilot was analysed to estimate and determine the appropriate number of test items for the second pilot, ensuring that each test could be completed within a single 40-minute lesson.

Overview

Table 3 presents the average time students in all age groups spent completing the AT test, which consisted of 65 items for Age Group 1 and 77 items for Age Groups 2 and 3. The mean completion times were 28 minutes for Age Group 1, 33 minutes for Age Group 2, and 28 minutes for Age Group 3, with standard deviations of 18, 17, and 13 minutes, respectively. While students in Age Groups 1 and 3 spent a similar amount of time on the test despite the difference in the number of test items, students in Age Group 2 took slightly longer. The longest completion time recorded was 184 minutes in Age Group 3, followed by 166 minutes in Age Group 2 and 141 minutes in Age Group 1. Figure 10 (available in a full version of the report) illustrates the distribution of students based on time spent on the AT test. Among the 2,715 students who participated in the first pilot, 23% (n = 613) spent more than 40 minutes completing the test, highlighting the need to reduce the number of test items. In contrast, only 30% (n = 828) of students completed the AT test within 18–30 minutes, whereas 55% (n = 1,840) completed the CT test within the same time frame.

Table 3. Mean, Standard Deviation, and Maximum Completion Time (in Minutes) for the AT Test Across Age Groups

(min)	Age Group 1	Age Group 2	Age Group 3	Age Groups 1–3
Mean completion time	28.16	33.44	27.89	30.16
SD of Completion time	18.29	16.86	13.26	16.72
Maximum completion time	140.99	165.63	183.73	183.73

Distribution of Time Spent on Each Section

In the first pilot, the AT test was designed to assess various AT sub-skills across six distinct sections. Each section contained a different number of items, adjusted for the three age groups to align with their developmental levels and the learning objectives outlined in their school curricula. The sections were structured as follows:

- **Section 1: Generalised Arithmetic**– This section included 23 items for Age Group 1, 25 items for Age Group 2, and 25 items for Age Group 3. It assessed students’ ability to recognise fundamental arithmetic relationships and mathematical properties and to extend them to more abstract generalisations.
- **Section 2: Equations and Inequalities** – This section contained 19 items for Age Group 1, 21 items for Age Group 2, and 22 items for Age Group 3. It focused on students’ understanding of the equal sign relationally and their ability to work with both equations and inequalities.
- **Section 3: Functional Thinking** – With 9 items for Age Group 1, 10 items for Age Group 2, and 11 items for Age Group 3, this section assessed students' ability to identify, generalise, and describe patterns. It also examined their understanding of relationships between co-varying quantities.
- **Section 4: Representation** – This section contained 3 items for Age Group 1, 4 items for Age Group 2, and 5 items for Age Group 3. It evaluated students' ability to use different forms of representation, such as diagrams, tables, symbols, and verbal descriptions, to organise information and develop a deeper understanding of mathematical relationships.
- **Section 5: Transformation** – Designed to assess students' ability to manipulate and restructure algebraic expressions while maintaining equivalence, this section included 5 items for Age Group 1, 8 items for Age Group 2, and 10 items for Age Group 3.
- **Section 6: Transversal Skills Needed for AT** – This final section measured a broad set of higher-order cognitive abilities necessary for algebraic thinking. It contained 7 items for Age Group 1, 8 items for Age Group 2, and 6 items for Age Group 3.

A detailed breakdown of these sections, including their specific objectives and item design, can be found in Report 3.2. The following analysis explores how students allocated their time across these sections, providing insights into their engagement with different AT sub-skills.

Table 4 shows that students in all age groups spent more time on Sections 1–3, with average times of 6–9 minutes, 7–10 minutes, and 7 minutes for Age Groups 1, 2, and 3, respectively. This is reasonable, given that these sections contained a significantly larger number of test items (36–48% of all test items) compared to the other sections. The number of students who spent excessive time on each section was also investigated. To ensure that all students can complete the assessment comfortably within a standard 40-minute lesson for future classroom implementation, we proposed maximum time limits for each section, based on the number of items assessing that particular AT sub-skill. A time limit of a maximum of 10 minutes was set for Sections 1–3, and a maximum of 5 minutes for Sections 4–6. Figures 11–16 (available in a full version of the report) illustrate the distribution of students based on the time spent on each section, considering the proposed maximum time limits.

Table 4. Mean Completion Time (in Minutes) for Each Section of the AT Test Across Age Groups

Mean completion time (min)	Age Group 1	Age Group 2	Age Group 3	Age Groups 1–3
1. Generalised arithmetic	5.94	7.43	6.69	6.72
2. Equations and inequalities	8.59	8.16	6.97	7.99
3. Functional thinking	7.23	10.47	7.36	8.53
4. Representation	1.24	0.98	1.66	1.25
5. Transformation	1.95	3.10	2.54	2.56
6. Transversal skills for AT	3.20	3.31	2.66	3.10

We further examined the amount of time students (N = 2,715) spent on six sections of the AT test. The data were analysed using the median, interquartile range (IQR, 25th–75th percentile), and 10th–90th percentile range to better understand the distribution of time spent on each section. The results for each section are as follows:

Section 1: Generalised Arithmetic

The **median** time spent on Section 1 was 6–7 minutes, indicating that half of the students completed this section in less time, while the other half took longer. Notably, the median was equal to the mean, suggesting a relatively symmetric distribution of time spent.

The **IQR** ranged from 3–4 minutes (25th percentile) to 9–10 minutes (75th percentile), showing that most students completed this section within 3–10 minutes. This is relatively long, given that the intended time limit for the section is 10 minutes.

The **10th–90th percentile range** extended from 1–2 minutes to more than 10 minutes, demonstrating broad variability in completion time, with some students taking significantly shorter or longer times. This suggests that while most students finished within the intended 10-minute time limit, some required additional time to complete this section.

Section 2: Equations and Inequalities

The **median** completion time was 7–8 minutes, meaning that half of the students finished within this range, while the other half took longer. The mean, however, was slightly higher (7–9 minutes), indicating a mildly skewed distribution.

The **IQR** ranged from 4–5 minutes to more than 10 minutes. This shows that most students took longer than expected, given the intended 10-minute time limit.

The **10th–90th percentile range** spanned from 1–2 minutes to over 10 minutes, reflecting significant variability in time spent on this section. This suggests that while most students finished within the intended 10-minute time limit, some required additional time to complete this section.

Section 3: Functional Thinking

The **median** time spent was 6–7 minutes, with a mean of 7–10 minutes, suggesting a right-skewed distribution. The mean, however, was 7–10 minutes, suggesting a skewed distribution to the right side.

The **IQR** ranged from 3–4 minutes to more than 10 minutes. This indicates that most students took longer than the intended 10-minute time limit for this section.

Similar to Section 2, the **10th–90th percentile range** extended from 1–2 minutes to more than 10 minutes, indicating a broad spread in completion time. This suggests that while most students finished within the intended 10-minute time limit, some required additional time to complete this section.

Section 4: Representation

The **median** time spent was 0–1 minute, suggesting that a substantial number of students completed this section very quickly. The mean was slightly higher (1–2 minutes), indicating a mild right-skewed distribution.

The **IQR** ranged from 0–1 minute to 1–2 minutes, showing that most students completed this section within 2 minutes. Given that this section contained only 3–5 items, this result is expected.

The **10th–90th percentile range** extended from 0–1 minute to 2–3 minutes, with most students finishing within 3 minutes. This is a favourable outcome since the intended time limit for this section is a maximum of 5 minutes.

Section 5: Transformation

The **median** time spent was 2–3 minutes. Notably, the mean matched the median value, indicating a symmetric distribution of time spent.

The **IQR** ranged from 0–1 minute to 3–4 minutes, meaning most students completed this section within 4 minutes, which is reasonable given the 5–10 items included in this section.

The **10th–90th percentile range** extended from 0–1 minute to more than 5 minutes. This suggests that while most students finished within the intended 5-minute time limit, some required additional time to complete this section.

Section 6: Transversal Skills Needed for AT

The **median** time spent was 2–3 minutes, with a mean of around 3 minutes.

Similar to Section 5, the **IQR** ranged from 1–2 minutes to 3–4 minutes, showing that most students completed this section within 4 minutes, with a concentration around 2–3 minutes. This indicates that most students spent rather short amounts of time on this section, given that there were 6–10 items.

The **10th–90th percentile range** spanned from 0–1 minute to more than 5 minutes, highlighting variability in completion times. While most students finished within the intended 5-minute time limit, some took significantly longer.

Summary

Overall, students spent relatively short durations on Sections 4, 5, and 6 (around 3 minutes), while Sections 1, 2, and 3 required moderate durations (approximately 6–8 minutes). The median time spent across sections ranged from 0–1 minute (Section 4) to 7–8 minutes (Section 2). The similarity between the mean and median in most sections suggests a generally balanced distribution, except for Sections 2 and 4, which showed mild right skewness. Notably, Section 3 exhibited the most pronounced skewed distribution, with a median of 6–7 minutes and a mean of 7–10 minutes.

The IQR analysis reinforced these findings, confirming that Sections 4, 5, and 6 required less time compared to Sections 1, 2, and 3. This was expected, as the first three sections contained more test items. However, variability in completion times differed across sections. Sections 1, 2, and 3 displayed considerable variability, with some students taking significantly longer or shorter times. In contrast, Sections 5 and 6 had moderate variability, while Section 4 showed the least variability, indicating a more consistent time distribution. Despite these differences, overall time spent remained concentrated around the intended time limits.

The 10th–90th percentile analysis highlighted the presence of outliers who either completed each section much faster or exceeded the expected completion time. Section 4 was the only exception, as only a small number of students exceeded the intended 5-minute time limit.

Recommendations for Test Revision

In the first pilot, the AT test comprised 66–79 items, leading to considerable variability in completion times across age groups. This poses challenges for future classroom implementation, as some students may require significantly more time to complete the test. To enhance feasibility, we decided to substantially reduce the number of test items, particularly in Sections 1, 2, and 3. To balance the assessment of AT skills with time constraints, we plan to increase the number of items in Section 4, as most students completed this section within 3 minutes—well below the intended 5-minute limit. The aim was to ensure that all students can complete the assessment comfortably within a standard 40-minute lesson for future classroom implementation, making it more practical for integration into regular classroom activities. The final number of items will be determined based on detailed statistical analyses, specifically, Confirmatory Factor Analysis (CFA) and Item Response Theory (IRT) , to ensure the test remains reliable and effective.

3.2 Quality Analysis of Test Items

Overview

A **two-parameter Item Response Theory (IRT) analysis** was carried out to assess the quality of the AT test items used in the first pilot (for detailed statistical results, see Report 3.3). The IRT provided insights into *discrimination* and the *difficulty* of each test item, guiding the selection and revisions of AT items in the same way as described in Section 2.2. The AT test was developed based on profound theoretical frameworks derived from the systematic literature review of existing assessment instruments and test items (see Report 3.1). The test was divided into six sections assessing different AT skills. Most of the sections were also divided into subsections to assess subskills. As a result, the test contained a large number of items. This made it challenging to carry out a **Confirmatory Factor Analysis (CFA)** to determine the relationships (correlations) among items and subitems. Therefore, no CFA was performed at this stage.

Based on the test completion time analysis, the number of items, particularly in Sections 1, 2, and 3, needs to be reduced substantially to enhance the feasibility of future classroom implementation. At the same time, the test also needs to contain a sufficient number of items to assess six AT skills and their subskills. The selection and revision of AT items were carried out to balance these two requirements.

First, all items with negative discrimination values were removed. Then the remaining items that had very low (below 0.65) or very high (above 4) discrimination values, as well as were extremely easy (item difficulty below -3) or extremely difficult (item difficulty above 3), were marked for more attention during the item selection process. Then these items were positioned along the ability scale of each test subsection for each age group based on their difficulty values. Finally, the items to be included in each subsection for the second pilot were selected to balance between the even distribution of difficulty levels and the AT skills to be assessed. If it was not necessary, the items marked for caution were not included in the test. However, if these items were included, their contents were revised by subject-matter experts to improve their discrimination and/or difficulty values. For example, the instruction of potentially problematic items was revised for better clarity; math expressions or the number of multiple choices were revised to decrease or increase the item difficulty. When appropriate, items that assessed the same AT skills of other age groups were included in the test for another age group.

Results and Actions Taken

Section 1: Generalised Arithmetic

Age Group 1 (Students Aged 9–10). Only one item (AT1_1.2B.2) demonstrated a negative discrimination value and was removed from the test. Several other items had either low or extremely high discrimination values. The item difficulty distribution for Section 1.1—the largest subsection—was skewed towards higher difficulty levels. Many items were identified as either too easy or too difficult. Following a detailed analysis of each item's content and IRT results, the number of items in this section was reduced from 23 in the first pilot to 17 in the second pilot.

Section 1.1 now includes 11 items, down from 15 in the first pilot. All selected items were converted from true-false to three-option multiple-choice format to reduce the likelihood of guessing and enhance test reliability. Despite low (AT1_1.1B.2) or high discrimination values (AT1_1.1A.8 and AT1_1.1B.10), these items were retained. Five items (AT1_1.1A.2, AT1_1.1A.8, AT1_1.1B.9, AT1_1.1B.10, and AT1_1.1A.13) were revised to improve the item discrimination and/or item difficulty distribution.

Section 1.2 remained at two items. AT1_1.2A.1 was carried over from the first pilot, while a revised version of AT2_1.2A.4—originally part of Age Group 2—was newly included.

Section 1.3 was reduced from four to two items (AT1_1.3A.1 and AT1_1.3A.2), both of which were revised to lower their difficulty levels.

Section 1.4 retained its two items (AT1_1.4B.1 and AT1_1.4B.2). Although both showed low discrimination and were challenging for this age group, they were revised from open-response to multiple-choice format to lower their difficulty levels.

Age Group 2 (Students Aged 11–12). No items in this group showed negative discrimination values. Only three items had discrimination values below 0.65, while seven had values above 4. The item difficulty distribution in Section 1.1 was mostly within the moderate (-1 to 1) and difficult (1 to 2) ranges. One item (AT2_1.4A.2) was extremely difficult, and two (AT2_1.1A.1 and AT2_1.1A.4) were too easy. These items were excluded. No items fell into the easy range (-2 to -1). Following careful evaluation, the total number of items in this section was reduced from 25 to 19.

Section 1.1 now contains 11 items, down from 15. All were converted from true-false to three-option multiple-choice items to reduce guessing and improve reliability. No items with problematic difficulty or discrimination values were retained. Four items (AT2_1.1B.7, AT2_1.1B.9, AT2_1.1B.10, and AT2_1.1A.15) were revised to improve the difficulty distribution.

Section 1.2 now includes three items (AT2_1.2A.1, AT2_1.2A.2, and AT2_1.2A.4), reduced from four. The content of these items was revised to lower their difficulty levels.

Section 1.3 contains three items, down from four. Two existing items (AT1_1.3A.1 and AT1_1.3A.2) were revised by reducing the number of answer choices from five to four, thereby lowering difficulty. A new item (AT1_1.3A.1) from Age Group 1 was added.

Section 1.4 remained unchanged in terms of item count, with two items (AT2_1.4B.1 and AT2_1.4B.2). To better match the difficulty level to this age group, both items were converted from open-response to multiple-choice format.

Age Group 3 (Students Aged 13–14). One item (AT3_1.2B.2) had a negative discrimination value, and another (AT3_1.1A.7) had a low value; both were removed. Seven items had extremely high discrimination values (above 4) and were also excluded. The item difficulty distribution for Section 1.1 was relatively balanced within the range of -2 to 2. No items had extreme difficulty values below -4 or above 4. Two items (AT3_1.1A.1 and AT3_1.2B.2) fell outside the acceptable range (-2 to 2) and were excluded. After content and IRT analysis, the total number of items in this section was reduced from 25 to 19. No items with inappropriate discrimination or difficulty values were retained.

Section 1.1 now comprises 12 items (down from 16). All were reformatted from true-false to a three-option multiple-choice format. Seven items (AT3_1.1A.3, AT3_1.1B.7, AT3_1.1B.8, AT3_1.1B.10, AT3_1.1A.11, AT3_1.1B.14, and AT3_1.1A.15) were revised to enhance the difficulty distribution.

Section 1.2 includes two items (formerly three). AT3_1.2A.1 was revised (increased from two to three answer choices) to raise difficulty, while AT3_1.2A.3 was unchanged.

Section 1.3 was reduced to two items. AT3_1.3A.1 was changed from open-response to four-option multiple-choice to lower difficulty. The other item (AT3_1.2A.3) remained unchanged.

Section 1.4 retained its three items. All were converted from open-response to four-option multiple-choice. Two items (AT3_1.4B.1 and AT3_1.4B.2) were carried over from the first pilot. One additional item (AT1_1.4B.1/AT2_1.4B.1), previously used with Age Groups 1 and 2, was added.

Section 2: Equations and Inequalities

Age Group 1 (Students Aged 9–10). No items showed negative discrimination values. Only two items (AT123_2.1A.1 and AT123_2.1B.1) had discrimination values between 0 and 0.65, and one (AT1_2.4A.5) above 4; all were removed. The item difficulty distribution for this section was skewed towards easier difficulty levels. There was no item covering the range of 1 and 2 (difficult level). Two items were relatively difficult (above 3), and one of them was removed.

Following a detailed analysis of each item's content and IRT results, the number of items in this section was reduced from 19 in the first pilot to 14 in the second pilot.

Section 2.1 now includes two items, down from three in the first pilot. Despite its notably high difficulty level (above 5), AT123_2.1A.2 was retained, as it was an anchor item across all age groups. AT123_2.1A.3 was also carried over from the first pilot.

Section 2.2 was reduced from seven to four items, three of which (AT1_2.2B.1, AT1_2.2A.2, and AT1_2.2A.6) were carried over from the first pilot. Only AT1_2.2B.5 was revised to increase its difficulty level.

Section 2.3 remained at two items. One (AT12_2.3A.2) was carried over from the first pilot, while the other one (AT12_2.3A.1) was revised to lower its difficulty level.

Section 2.4 was reduced from five to four items, all of which (AT1_2.4A.2, AT1_2.4B.3, AT1_2.4A.4, and AT1_2.4B.5) were carried over from the first pilot.

Section 2.5 remained at two items. One (AT1_2.5A.2) was carried over from the first pilot, while the other one (AT1_2.5A.1) was revised to lower its difficulty level.

Age Group 2 (Students Aged 11–12). One item (AT123_2.1B.2) had a negative discrimination value, while two (AT123_2.1A.1 and AT123_2.1B.1) had low discrimination values between 0 and 0.65; all were removed. No items had discrimination values above 4. The item difficulty distribution for this section was heavily skewed toward easier items. There was no item covering the range of 0.5 and 2 (moderately difficult and difficult levels). Four items (AT123_2.1B.2, AT2_2.2B.1, AT2_2.2A.3, and AT2_2.2A.7) were too easy (below -2); all were excluded. Following a detailed analysis of each item's content and IRT results, the number of items in this section was reduced from 21 in the first pilot to 14 in the second pilot.

Section 2.1 was reduced from three to two items, which are the same anchor items as for Age Group 1. However, AT123_2.1A.2 was not extremely difficult for this age group (approximately 2.4) compared to the younger group.

Section 2.2 was reduced from eight to four items, three of which (AT2_2.2A.1, AT2_2.2B.4, and AT2_2.2B.6) were carried over from the first pilot. Only AT2_2.2B.5 was revised to increase its difficulty level.

Section 2.3 remained at two items. One (AT12_2.3A.2) was carried over from the first pilot, while the other one (AT 2_2.3 New) was newly added to balance the item difficulty distribution.

Section 2.4 was reduced from six to four items. Two items (AT2_2.4B.3 and AT2_2.4A.4) were carried over from the first pilot, while the other two, previously used with Age Group 3 (AT3_2.4A.6 and AT3_2.4A.8), were added.

Section 2.5 remained at two items. One (AT2_2.5A.1) was carried over from the first pilot, while the other one (AT2_2.5A.2) was revised to lower its difficulty level.

Age Group 3 (Students Aged 13–14). No items had a negative discrimination value. Some items had discrimination values between 0 and 0.65, demonstrating a limited ability to differentiate between students of varying ability levels. Most of these items were removed. AT123_2.1A.2 was the only item with a discrimination exceeding 4. The item difficulty distribution for this section appears to most balanced compared to the younger groups. However, the distribution was slightly skewed toward easier items, and there was no item covering the range of 0.5 and 1.5 (moderately difficult and difficult levels). Following a detailed analysis of each item's content and IRT results, the number of items in this section was reduced from 22 in the first pilot to 15 in the second pilot.

Section 2.1 now includes two items, down from three in the first pilot. Despite their low (AT123_2.1A.3) and high discrimination values (AT123_2.1A.2), both items were retained as anchor items across all age groups.

Section 2.2 was reduced from five to three items, two of which (AT3_2.2A.1 and AT3_2.2B.3) were carried over from the first pilot. Only AT3_2.2A.4 was revised to increase its difficulty level.

There was no Section 2.3 (Work with picture variables) for this age group.

Section 2.4 was reduced from 11 to six items, five of which (AT3_2.4B.3, AT3_2.4A.4, AT3_2.4A.6, AT3_2.4A.8, and AT3_2.4A.11) were carried over from the first pilot. Only AT3_2.4A.9 was revised to decrease its difficulty level.

Section 2.5 increased from three to four items. Two items (AT3_2.5A.1 and AT3_2.5A.2) were carried over from the first pilot. AT3_2.5B.3 was revised to lower its difficulty level, while AT3_2.5.4 New was newly added to the test.

Section 3: Functional Thinking

Age Group 1 (Students Aged 9–10). All items in this group had discrimination values exceeding 0.65, indicating effective differentiation between students of varying ability levels. AT1_3.2A.2 was the only item with a discrimination value above 4, thus removed. The item difficulty distribution for this section appeared well-balanced but slightly skewed toward more difficult items. All items were in the normal range of difficulty between -2 and 2. Only AT1_3.3A.2 had a difficulty value exceeding 2 and was excluded. There was no item covering the range of -1.5 and -2 (easy level). Following a detailed analysis of each item's content and IRT results, the number of items in this section was reduced from nine in the first pilot to seven in the second pilot.

Section 3.1 now comprises three items (down from four). Two items (AT12_3.1B.2 and AT1_3.11B.1) were carried over from the first pilot. One additional item (AT1_3.1New) was added to this section.

Section 3.2 includes two items (formerly three). Both items (AT1_3.2B.1 and AT1_3.2A.3) were carried over from the first pilot.

Section 3.3 retained its two items. While AT1_3.3B.1 remained unchanged, AT1_3.3B.2 was revised to reduce its difficulty.

Age Group 2 (Students Aged 11–12). All items in this group had discrimination values within the range of 0.65 and 4, indicating effective differentiation between students of varying ability levels. All items were in the normal range of difficulty between -2 and 2. The difficulty distribution for this section is slightly skewed toward easy items. There was no item covering the range of 1.5 and 2 (difficult level). There were also gaps around difficulty values of -1.5 and -1, as well as 0.1 and 0.5. Following a detailed analysis of each item's content and IRT results, the number of items in this section was reduced from 10 in the first pilot to seven in the second pilot.

Section 3.1 now comprises three items (down from four). All three items (AT23_3.1A.1, AT2_3.11B.2, and AT2_3.11B.3) were carried over from the first pilot.

Section 3.2 includes two items (formerly three). AT2_3.2A.2 was carried over from the first pilot, while AT2_3.2B.3 was revised to decrease its difficulty as it acts as an anchor item for Age Groups 2 and 3.

Section 3.3 now comprises two items (down from three). While AT2_3.3B.1 remained unchanged, AT2_3.3B.3 was revised to reduce its difficulty as it acts as an anchor item across aged groups.

Age Group 3 (Students Aged 13–14). All items in this group had discrimination parameter values exceeding 0.65, indicating they effectively differentiated between students of varying ability levels. There were no items with discrimination estimates above 4, ensuring a stable range without extreme variations. The distribution of difficulty parameter estimates appears to be well-balanced. All items were in the normal range of difficulty between -2 and 2. There was no item covering very easy (-2) and difficult levels (2). There were also difficulty gaps between -1.5 and -1 as well as 0.1 and 0.5. Following a detailed analysis of each item's content and IRT results, the number of items in this section was reduced from nine in the first pilot to seven in the second pilot.

Section 3.1 now comprises two items (down from four). AT3_3.1B.3 was carried over from the first pilot, while AT3_3.1B.2 was revised to decrease its difficulty.

Similar to the first pilot, Section 3.2 now comprises three items. AT3_3.2B.1 was revised to increase its difficulty as an anchor item across age groups. AT3_3.2B.2, an anchor item for Age Groups 2 and 3, was also revised to decrease its difficulty. AT3_3.2.3 New was additionally added to this section as a difficult item.

Section 3.3 now comprises two items similar to the first pilot. While AT3_3.3A.1 remained unchanged, AT3_3.3 New was additionally added as an anchor item across aged groups.

Section 4: Representation

Age Group 1 (Students Aged 9–10). All items had positive discrimination values exceeding 0.65, suggesting that they were generally effective in differentiating students based on their abilities. Two items had discrimination values below 4. Four items had discrimination values above 4, two of which had extremely high discrimination values and were removed. The item difficulty distribution for this section was skewed towards more difficult levels. All items had difficulty parameter values ranging between 0.5 and 2, meaning none were classified as particularly easy or overly difficult. Following a detailed analysis of each item's content and IRT results, the number of items in this section was increased from three in the first pilot to five in the second pilot.

Section 4.1 was increased from two to three items, which are the same anchor items as for Age Group 1. Two of them (AT12_4.1A.1 and AT12_4.1A.2) were carried over from the first pilot, while AT12_4.1A.1 New was newly added to balance the item difficulty distribution.

Section 4.2 was increased from one to two items. Both items (AT1_4.2.1 and AT1_4.2.2) were newly added as anchor items across age groups.

Age Group 2 (Students Aged 11–12). All items had positive discrimination values. AT2_4.2B.1, the only item with a discrimination value below 0.65 (as well as a difficulty value above 2) was removed. Three items had discrimination values above 4, two of which, with extremely high discrimination values, were also removed. The item difficulty distribution for this section was skewed towards moderate and difficult levels. All items had difficulty parameter values ranging between -0.2 and 1.3, indicating a lack of easier test items. Following a detailed analysis of each item's content and IRT results, the number of items in this section was increased from four in the first pilot to seven in the second pilot.

Section 4.1 was increased from two to three items, which are the same anchor items as for other age groups. One item (AT12_4.1A.1) was carried over from the first pilot, while the other two (AT12_4.1A.1 New and AT2_4.1.3), anchor items for other age groups, were newly added to balance the item difficulty distribution.

Section 4.2 was increased from two to four items. Two items (AT2_4.2A.1 and AT2_4.2A.2) were carried over from the first pilot, while the other two (AT2_4.2A.1 New and AT2_4.2A.2

New), anchor items for other age groups, were newly added to balance the item difficulty distribution.

Age Group 3 (Students Aged 13–14). All items had positive discrimination values exceeding the threshold of 0.65, suggesting that they were generally effective in differentiating students based on their abilities. Three of four items with discrimination values exceeding 4 were removed. The item difficulty distribution for this section was skewed towards moderate and more difficult levels. All items had difficulty parameter values ranging between -0.15 and 1.5, indicating a lack of easier test items. Following a detailed analysis of each item's content and IRT results, the number of items in this section was increased from five in the first pilot to seven in the second pilot.

Section 4.1 was increased from two to three items. One item (AT3_4.1A.2) was carried over from the first pilot, while the other two (AT3_4.1A.1 New and AT3_4.1A.2 New) were newly added to balance the item difficulty distribution and as anchor items across age groups.

Section 4.2 was increased from three to four items. One item (AT3_4.2A.3) was carried over from the first pilot. The other three were newly added to balance the item difficulty distribution: two (AT3_4.2A.1 New and AT3_4.2A.3 New) as anchor items across age groups, and one (AT3_4.2A.2 New) only for this age group.

Section 5: Transformation

Age Group 1 (Students Aged 9–10). All items had discrimination values exceeding the threshold of 0.65, suggesting that they were generally effective in differentiating students based on their abilities. Four items had discrimination values above 4, three of which were removed. The item difficulty distribution for this section was skewed towards more difficult levels. All items had difficulty parameter values ranging between 0.3 and 2, suggesting that students generally found these subitems difficult. Following a detailed analysis of each item's content and IRT results, the number of items in this section remained at five items as in the first pilot.

Section 5.1 remained as three items. Two items (AT1_5.1B.2 and AT1_5.1B.3) were carried over from the first pilot, which are the same anchor items in other age groups. The other item (AT1_5.1B.1 New) was newly added to balance the item difficulty distribution.

Section 5.2 remained as two items. Both items (AT1_5.3B.1 and AT1_5.3A.2) were carried over from the first pilot as anchor items across age groups.

Age Group 2 (Students Aged 11–12). All items had positive discrimination values. AT2_5.3B.4, the only item with a discrimination value below 0.65, was removed. Six items had discrimination values above 4, only one of which was retained. The item difficulty distribution for this section was skewed towards moderate and more difficult levels. Most items had difficulty parameter values ranging between 0.1 and 1.9; only one item (AT2_5.1B.1) had a negative difficulty value, meaning that most of the items were relatively challenging for students in this age group. Following a detailed analysis of each item's content and IRT results, the number of items in this section was reduced from eight in the first pilot to five in the second.

Section 5.1 was reduced from four to three items. AT2_5.1B.2 was carried over from the first pilot; AT2_5.1B.3 was revised to decrease its difficulty level; AT2_5.1B.1 New was newly added to balance the item difficulty distribution.

Section 5.2 was decreased from four to two items. Both items (AT2_5.3B.1 and AT2_5.3A.2) were carried over from the first pilot.

Age Group 3 (Students Aged 13–14). All items had positive discrimination values exceeding 0.65, suggesting that they were generally effective in differentiating students based on their abilities. Only four of the 20 subitems had discrimination parameter estimates exceeding 4 and thus were removed. As shown in All difficulty parameter values fell within the narrow range of -0.4 to 0.5, this section was moderately difficult for students in this age group. Following a detailed analysis of each item's content and IRT results, the number of items in this section was reduced from 10 in the first pilot to eight in the second pilot.

Section 5.1 remained as three items, all of which are the same anchor items as for other age groups. One item (AT3_5.1B.1) was carried over from the first pilot, while the other two (AT3_5.1B.1 New and AT3_5.1B.3 New) were newly added to balance the item difficulty distribution.

Section 5.2 was reduced from seven to five items. AT3_5.3B.4 and AT3_5.3B.7 were carried over from the first pilot; AT3_5.3A.5 was revised to increase its difficulty level; AT3_5.2.1 New (previously used with Age Groups 1 and 2) and AT3_5.2.5 New were added.

Section 6: Transversal Skills Needed for AT

Age Group 1 (Students Aged 9–10). Only one item had a negative discrimination value and thus was removed. Among the remaining subitems, two items had discrimination values below 0.65, and one had a discrimination value exceeding 4. All three items were also removed. As shown, the distribution of difficulty values was relatively balanced, covering a reasonable range from easy to difficult (between -2 and 2). Only two items fell outside this range (too easy or too difficult for this age group) and thus were removed. Overall, the item difficulty distribution was skewed towards easier levels. Following a detailed analysis of each item's content and IRT results, the number of items in this section was reduced from seven to six items.

Three items were carried over from the first pilot, two of which (AT12_6.1A.1 and AT12_6.1A.2) are the same anchor items in other age groups, while AT1_6.5A.1 was only for this age group. Two items were revised: AT12_6.3B.1 to reduce its difficulty and AT1_6.4B.1

to increase its difficulty. AT1_6.2 New was newly added to balance the item difficulty distribution.

Age Group 2 (Students Aged 11–12). The item with a negative discrimination value and the item with a discrimination value exceeding 4 were removed. Three items had discrimination values below 0.65, two of which were also removed. The item difficulty distribution for this section was well-balanced, suggesting that this test evaluated transversal skills of students with different proficiency levels. Only three items had difficulty values outside the range between -2 and 2 and thus were removed. Following a detailed analysis of each item's content and IRT results, the number of items in this section was reduced from eight to seven items.

Four items (AT12_6.1A.1, AT23_6.2A.1, AT23_6.5A.1, and AT2_6.4A.1) were carried over from the first pilot as anchor items across age groups. Two items were revised: AT12_6.3B.1 to increase its difficulty and AT23_6.3B.1 to improve its clarity, thus improving its discrimination value. AT2_6.4 New was newly added to balance the item difficulty distribution.

Age Group 3 (Students Aged 13–14). The item with a negative discrimination value and the item with a discrimination value exceeding 4 were removed. Four items had discrimination values below 0.65, three of which were also removed. The item difficulty distribution for this section was skewed towards more difficult levels. Almost all items had difficulty parameter values within the range of -2 and 2. Following a detailed analysis of each item's content and IRT results, the number of items in this section was increased from six to seven items.

Five items were carried over from the first pilot: three (AT23_6.2A.1, AT3_6.4A.1, and AT23_6.5A.1) as anchor items across age groups, two (AT23_6.4A.1 and AT3_6.4A.2) only used for this age group. One item (AT23_6.3B.1) was revised to improve its clarity, thus improving its discrimination value. AT123_6.1 New, used also for Age Groups 1 and 2, was newly added to balance the item difficulty distribution.

4. COMATH 1–3 for the Second Pilot

Unlike the first pilot, there was only one version of the CT and AT test. To ensure that all students can complete the assessment comfortably within a standard 40-minute lesson for future classroom implementation, we reduced the total number of items. Additionally, a series of anchor items was included across two or all three levels of COMATH to maintain continuity and enable the comparison of students' skills across different age groups.

4.1 CT Test

The CT test for the second pilot contains 14 items for each age group. Each section now consists of seven items, maintaining a balanced assessment of CT skills while improving practicality for classroom use. Table 5 provides all CT test items included in COMATH 1–3 for the second pilot. In this final version, we compiled a total of **24 CT test items**.

4.2 AT Test

Overall, we dramatically decreased the total number of AT test items. Table 6 compares the number of AT items included in COMATH 1–3 for the first and second pilots. In the full version of the report, we present examples of AT test items included in COMATH 1–3 for the second pilot.

Table 6. Number of AT test items included in COMATH 1–3 for the second pilot

	COMATH 1 (9–10 y)		COMATH 2 (11–12 y)		COMATH 3 (13–14 y)	
	Pilot 1	Pilot 2	Pilot 1	Pilot 2	Pilot 1	Pilot 2
Section 1	23	17	25	19	25	19
Section 2	19	14	21	14	22	15
Section 3	9	7	10	7	9	7
Section 4	3	5	4	7	5	7
Section 5	5	5	8	5	10	8
Section 6	7	6	8	7	6	7
Total	66	54	76	59	77	63

References

Baker, F. B. (2001). *The basics of item response theory*. Retrieved from http://ericae.net/irt/baker.

Gyamfi, A., & Acquaye, R. (2023). Parameters and models of item response theory (IRT): A review of literature. *Acta Educationis Generalis, 13(3),* 68–78.