Co-funded by
the European Union

Pilot 1 and feedback gathering. Summary

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 Esa Satomaa under WP3.

Revised by Valentina Dagienė

Table of Contents_Toc207368017

Co-funded by the European Union

Pilot 1 Computational Thinking (CT) test

Pilot 1 CT test data

The collected data comprised answers and response times from the Pilot 1 CT test, taken by students from six different countries (Finland, Sweden, Lithuania, Turkey, Hungary, and Spain). There were three different test versions, each designed for a specific age group: COMATH1 for ages 9–10, COMATH2 for ages 11–12, and COMATH3 for ages 13–14. Each test item also had an A and B version, with each student completing one. The version of each test item was assigned randomly. Each test item fell into one of two subgroups aimed at assessing either algorithmic thinking skills (ALG) or algorithmic thinking along with other CT skills (OTH). After data processing the final dataset included 3350 students and 53268 valid (non-NA) responses, see the main Pilot 1 report for details on the data processing.

Pilot 1 CT methods

Each age group was analysed separately. Main analyses were first of all EFA and CFA (Exploratory Factor Analysis and Confirmatory Factor Analysis) to explore potential underlying factor structures in the test items. This was done to merged data where A and B versions of test items were considered the same test item because of technical statistical reasons. Given their near-identical nature, it was reasonable to assume that students would respond to A and B versions in a similar manner.

Second main analysis was IRT (Item Response Theory) analysis that was used to assess the difficulty and discrimination ability of the test items, with A and B versions analysed separately. A two-parameter IRT model was applied to estimate both a difficulty and a discrimination parameter for each item. Discrimination ability means test items ability to differentiate between students based on their ability.

Third main analysis was reliability analysis, this was also done to data where A and B versions were merged together for technical statistical reasons. This included assessing test reliability via Cronbach's alpha coefficient and 3-factor Omega coefficient, and assessing test item performance via item-total-correlations, which measure the correlation between each item's score and the total test score.

Pilot 1 CT results

Pilot 1 CT EFA summary

EFA did not reveal distinct factors separating algorithmic items from other items, as might have been expected. Instead, the most suitable factor structures—based on diagnostic tools, such as the Tucker-Lewis Index (TLI) and the Root Mean Square Error of Approximation (RMSEA)—were a three-factor structure for age groups 1 and 3 and 2-factor structure for age group 2 as the three-factor model included a factor with only one test item. These factors appeared to correspond to item difficulty levels (difficult, easy, and intermediate items). This type of factor structure is commonly observed when binary variables are analysed using standard factor analysis.

Co-funded by the European Union

Pilot 1 CT CFA summary

CFA was conducted to evaluate three models across all age groups: 1-factor model, 2-factor model (ALG and OTH test items) and a model with the factor structure identified by EFA. Model indicators like TLI and RMSEA and AIC values consistently supported the factor structures identified by EFA as the best-fitting models across all age groups. However, both the one-factor model (assuming a single underlying concept) and the two-factor model (distinguishing algorithmic and other items) also produced comparable diagnostic values, suggesting they remain plausible representations of the data. Chi-squared test supported (test p-value > 0.05) only the EFA based three-factor model for age group 1 (p = 0.062). The EFA based two-factor model for age group 2 had a p-value of 0.022 and all other models had p-values below 0.001. It is important to note that the chi-square test is highly sensitive to large sample sizes, meaning that low p-values may not necessarily indicate poor model fit.

The **exploratory and confirmatory factor analyses** used in this study assume that the observed variables are continuous. However, in this case, the variables are **binary** (correct/incorrect responses). Although binary variables are generally considered to reflect an underlying continuous CT skill, this assumption may affect the analysis. For example, the model may underestimate some test items' factor loadings, meaning their true contribution to the construct could be higher than indicated.

Pilot 1 CT IRT summary

IRT analysis revealed that the distribution of item difficulties was skewed towards higher difficulty levels, making the test less reliable for lower-performing students. This effect was most pronounced in Age Group 1, suggesting that the test was particularly challenging for younger participants. The distribution of difficulty parameter estimates was most evenly spread within the (-2, 2) range for Age Group 3, indicating that test items were best suited for this age group.

There were some test items that had problematic IRT parameters. Test item ALG-07-A had abnormally high difficulty parameter estimates (13.35 in age group 1 and 4.68 in age group 2) and low discrimination estimates (0.23 in age group 1 and 0.54 in age group 2), which indicates that this test item does not perform very well in either age group. Test item ALG-07-B had abnormally high difficulty parameter estimate in age group 2 (8.36) and abnormally high discrimination estimate in age group 1 (7.62). Test item OTH-07-A had abnormally high discrimination parameter estimate in age group 1 (9.47) as did test item OTH-07-B (5.45). Finally, test item ALG-05-A had abnormally high discrimination parameter estimate in age group 3 (4.6) and test item OTH-09-B had low discrimination parameter estimate in age group 3 (0.57).

Pilot 1 CT test reliability summary

Cronbach's alpha coefficients (raw alpha) for Pilot 1 CT test were 0.74 for age group 1, 0.71 for age group 2 and 0.78 for age group 3. Omega total coefficients with 3 factors were 0.78 for age group 1, 0.77 for age group 2 and 0.84 for age group 3. These coefficients indicate an adequate reliability in all age groups, especially considering that the test data is binary (1 for correct answer and 0 for incorrect answer) and Cronbach's alpha and Omega assume continuous variables. According to the coefficients reliability was best in age group 3 and least good in age group 2.

Some test items had low item-rest correlations (correlations between the item score and the total test score excluding the item in question), indicating low item performance. An item-rest correlation (hereafter RIR; referred to as *r.drop* in the main report) below 0.2 is considered low in the analysis. In age group 1, these test items were ALG-07 (RIR = 0.041), OTH-07 (RIR = 0.072), and OTH-08 (RIR = 0.192). In age group 2, they were ALG-03 (RIR = 0.17), ALG-07

Co-funded by the European Union

(RIR = 0.076), ALG-08 (RIR = 0.139), OTH-07 (RIR = 0.184), OTH-08 (RIR = 0.184), and OTH-14 (RIR = 0.162). In age group 3, they were ALG-05 (RIR = 0.17) and OTH-05-A (RIR = 0.17).

Pilot 1 CT problematic test item summary

ALG-07 (both A and B versions) and OTH-07 (both A and B versions) were considered significantly problematic test items because they had low correlations with other items, weak factor loadings, abnormal IRT parameter estimates and low item-rest-correlations. Test items ALG-08 (A and B versions), OTH-08, ALG-05 (A and B versions), OTH-05-A, ALG-03 and OTH-09-B were considered moderately problematic items because they had some of the mentioned or similar problems. Furthermore, some test items were considered possibly too difficult even though they did not show issues in other diagnostic measures. These were OTH-02 (difficulty parameter estimate = 2.69), ALG-10-A (difficulty parameter estimate = 2.85 for age group 1) and ALG-10-B (difficulty parameter estimate = 3.35 for age group 1).

Co-funded by the European Union

Pilot 1 Algebraic Thinking (AT) test

Pilot 1 AT test data

Pilot 1 AT test data was also from 6 countries, had 3 age groups and A and B versions from test items assigned randomly. Instead of 2 test subgroups it had 6: generalized arithmetic; equivalence, equations, and inequalities; functional thinking; representation; transformation; and transversal skills for AT. After data processing the final dataset included 2715 students and 132374 valid (non-NA) responses, see the main Pilot 1 report for details on the data processing.

Pilot 1 AT methods

The dataset comprised 400 distinct test subitems. Due to the large number of subitems, each combination of age group and test subgroup was analyzed separately. The methods used for the CT test could not be directly applied to the AT test. Since students completed only either the A-version or the B-version of a given subitem, it was not possible to calculate correlation coefficients between the two versions of the same test item. Moreover, merging the A and B versions of a subitem was not considered appropriate, as some versions had substantial differences. For these reasons, factor analysis and reliability analysis were not conducted for the AT test.

Pilot 1 AT results

Pilot 1 AT IRT summary: Generalised Arithmetic

There were test items with problematic IRT parameter estimates in all age groups. Overall, the generalised arithmetic section of the test appeared to function best for Age Group 3, compared to the younger age groups. This group had fewer extreme parameter estimates, and the difficulty estimates were distributed more evenly. Additionally, the number of test subitems was 46 for Age Group 1 and 50 for Age Groups 2 and 3, suggesting that the test structure remained consistent across these groups.

Pilot 1 AT IRT summary: Equivalence, Equations and Inequalities

Several test items shared across all three age groups, identified by item codes beginning with AT123, exhibited problematic parameter estimates in one or more age groups. For example, the subitems AT123_2.1A.1 and AT123_2.1B.1 demonstrated a consistently poor ability to differentiate between students in all age groups. However, the distribution of difficulty parameter estimates appeared to be most balanced in Age Group 3.

Pilot 1 AT IRT summary: Functional thinking

The functional thinking subitems showed stable parameter estimates across all three age groups, suggesting that the model effectively assessed students' abilities. Few subitems had problematic parameter estimates, indicating that most items reliably measured functional thinking skills without significant inconsistencies.

The distribution of difficulty parameter estimates was similar across age groups, with slight differences in the ranges covered. No subitems had difficulty estimates below -4 or above

4, confirming that all items fell within a reasonable difficulty range. Overall, the distributions appeared well-balanced. However, minor gaps at the lower and higher ends suggest that adding very easy or very difficult subitems could enhance the test's ability to assess the full spectrum of student proficiency.

Pilot 1 AT IRT summary: Representation

The range of difficulty parameter estimates for all representation subitems was relatively narrow across all age groups. Additionally, almost all difficulty estimates were positive, indicating that, according to the model, most test subitems were of above-average difficulty. This suggests that the test may not fully capture the abilities of lower-performing students, as there were few subitems at the easier end of the scale. Introducing additional low-difficulty items could improve the test's overall balance and effectiveness.

Another notable finding was that test item AT12_4.1 exhibited abnormally high discrimination parameter estimates in both Age Groups 1 and 2. While high discrimination values indicate that the item effectively differentiates between students of varying abilities, extremely large estimates may suggest potential model instability or overfitting. Further examination of this item may be necessary to determine whether it functions as intended within the assessment framework.

Pilot 1 AT IRT summary: Transformation

The analysis of transformation subitems across age groups revealed similarities in discrimination and difficulty parameter estimates for Age Groups 1 and 2.

No subitems had discrimination parameter estimates below 0, and nearly all had values above 0.65, confirming their reliability in distinguishing between students with different proficiency levels. However, abnormally high discrimination values were observed, particularly in Age Groups 1 and 2, with multiple subitems exceeding a threshold of 4. While such high values suggest strong differentiation, extreme estimates may indicate model instability or overfitting. In contrast, Age Group 3 displayed more moderate discrimination estimates, suggesting better statistical stability.

Difficulty estimates for Age Groups 1 and 2 were skewed toward the moderate-to-high range, indicating that most subitems were challenging, with few easier items available to assess lower-performing students. In Age Group 3, the difficulty range was slightly broader, covering both easier and more difficult items, but remained narrow (-0.5 to 0.5), limiting its ability to differentiate students effectively.

Addressing these issues would improve the test's effectiveness. High-discrimination items in Age Groups 1 and 2 should be reviewed to ensure stable measurement, and additional easier items should be introduced to enhance assessment balance. Expanding the difficulty range in Age Group 3 would allow for a more comprehensive evaluation of varying proficiency levels.

Pilot 1 AT IRT summary: Transversal Skills for AT

The IRT analysis across all three age groups revealed consistent patterns in discrimination and difficulty parameter estimates.

In all age groups, the discrimination parameter estimates confirm the reliability of most subitems in distinguishing between students of different proficiency levels. Only a few subitems had estimates below 0.65, indicating weak differentiation, while one or two subitems in each group exhibited abnormally high discrimination values exceeding 4. These extreme estimates, particularly in Age Groups 1 and 2, may suggest model instability or overfitting and should be interpreted with caution.

The distribution of difficulty parameter estimates for transversal skills subitems was generally well-balanced, with Age Group 2 displaying the most even spread. However, subitem AT23_6.4B.1 consistently demonstrated highly abnormal parameter estimates across Age Groups 2 and 3, including a negative discrimination value and an extremely low difficulty estimate. These irregularities suggest that the subitem may not function as intended and warrants further investigation or potential removal from the assessment.

Pilot 1 Student surveys

As part of this pilot study, students were asked to complete a digital survey immediately after taking the CT test. The survey comprised two sections: the first included two background questions (age and grade level) and one question on self-assessed mathematics skills, while the second contained six questions related to the CT test experience. A similar survey was administered following the AT test, with six questions specific to that assessment. Additionally, qualitative feedback was collected to refine and improve the assessment instrument through usability testing conducted in Turkey in April 2024.

In this summary, only mean scores and standard deviations that were calculated separately for self-assessed mathematics skills and the six survey statements across three age groups: under 11 years, 11–12 years, and over 12 years are reported. These statistics for the Pilot 1 CT test are presented in Table 1 and for the Pilot 1 AT test in Table 2. Survey questions were on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree).

Table 1. Mean scores and standard deviations for self-assessed mathematics skills and CT test survey responses across age groups.

	Age Group 1		Age Group 2		Age Group 3	
Statement	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
“I am good at math.”	3.76	1.12	3.51	1.08	3.30	1.11
“I did my best to complete the test.”	4.19	1.11	4.20	0.94	3.83	1.08
“The test was easy for me.”	3.16	1.13	3.20	1.01	3.07	1.03
“It was easy for me to complete the test in this digital system.”	3.47	1.23	3.62	1.08	3.58	1.19
“The test helped me to evaluate my skills.”	3.78	1.22	3.61	1.20	3.24	1.22
“It was nice to take the test.”	3.77	1.32	3.57	1.32	3.13	1.37
“I would like to take a similar test in the future.”	3.43	1.44	3.23	1.43	2.95	1.46

Table 2. Mean scores and standard deviations for self-assessed mathematics skills and AT test survey responses across age groups.

	Age Group 1		Age Group 2		Age Group 3	
Statement	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
“I did my best to complete the test.”	4.06	1.17	3.88	1.14	3.49	1.23
“The test was easy for me.”	2.93	1.23	3.01	1.13	2.87	1.20
“It was easy for me to complete the test in this digital system.”	3.33	1.31	3.49	1.20	3.30	1.25
“The test helped me to evaluate my skills.”	3.62	1.30	3.26	1.25	2.91	1.27
“It was nice to take the test.”	3.45	1.46	3.03	1.39	2.68	1.33
“I would like to take a similar test again in the future.”	3.14	1.54	2.88	1.48	2.58	1.41

The reliability of the survey responses was assessed using Cronbach’s alpha and omega total (3 factors), which measure internal consistency. For all participants and for each age group separately Cronbach’s alpha and Omega total indicated good reliability (coefficients > 0.8).

Co-funded by
the European Union

Pilot 1 Teacher survey

Following the completion of the AT and CT test pilots, teachers were invited to participate in an online survey designed to gather insights into their backgrounds, experiences, and perceptions of the tests. A total of 60 teachers participated in the survey. In addition to this survey, qualitative feedback was collected through a separate survey and a focus group interview conducted during a teacher workshop held in Turkey in April 2024.

In this summary only histograms of teachers answers to three survey questions, “The test measures AT/CT skills well” (Figure 1), “The difficulty of the AT/CT test was appropriate for my students” (Figure 2) and “The AT/CT test does not need any revision” (Figure 3) are reported.

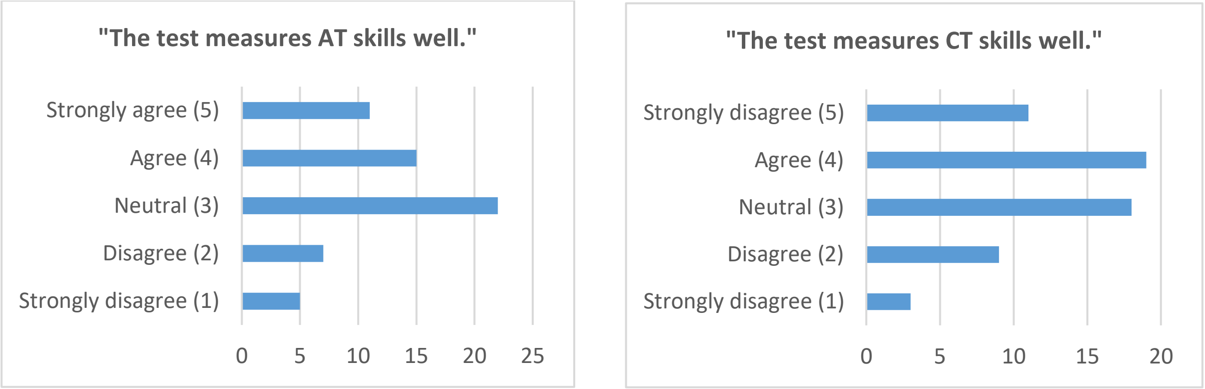

Figure 1. Teachers' perceptions of the validity of the AT test compared to the CT test.

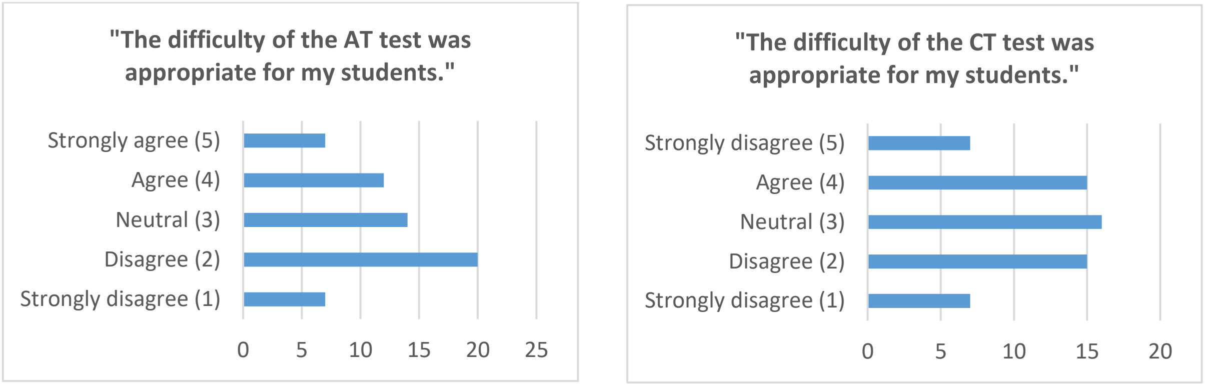

Figure 2. Teachers' perceptions of the appropriateness of the difficulty level of the AT test compared to the CT test.

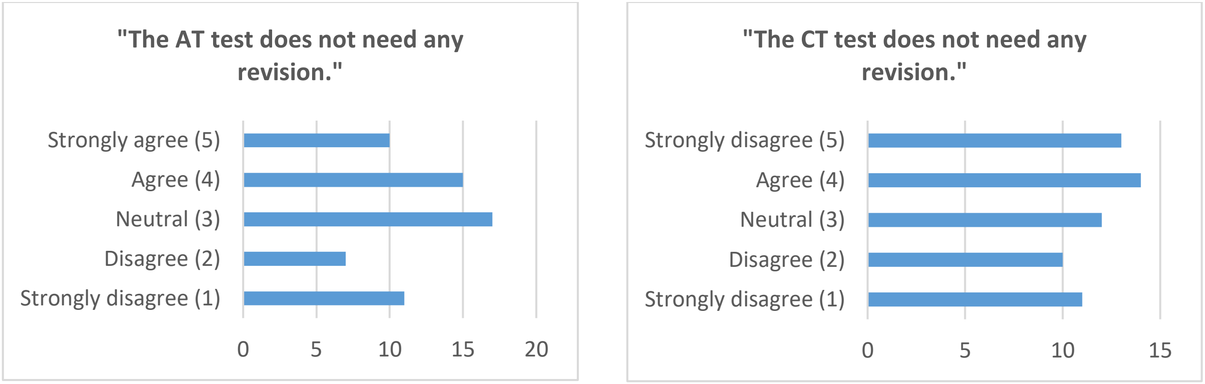

Figure 3. Teachers' perceptions of the need for revision of the AT test compared to the CT test.