Co-funded by
the European Union

Pilot 2 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 were developed by Daranee Lehtonen and Esa Satomaa under WP3.

Revised by Valentina Dagienė

Co-funded by
the European Union

Table of Contents_Toc207367744

Co-funded by the European Union

Pilot 2: Computational Thinking (CT) test

Pilot 2 CT test data

The data included answers and response times from the Pilot 2 CT test, taken by students from six countries: Finland, Sweden, Lithuania, Turkey, Hungary, and Spain. Three test versions targeted different age groups: COMATH1 (ages 9–10), COMATH2 (11–12), and COMATH3 (13–14). Unlike the Pilot 1 test, each item had only one version. Items were categorized into two subgroups: algorithmic thinking (ALG) or algorithmic thinking combined with other CT skills (OTH). Each test contained 14 items. After processing, the final dataset included 42630 non-NA answers from 3045 students. See the Pilot 2 main report for details on the data processing.

Pilot 2 CT methods

Each age group took a separate test and was analyzed independently. As this is a brief summary, methods are presented without the full details available in the main report.

Test reliability was tested by calculating Cronbach’s alpha coefficient and deflation-corrected alpha coefficient, which is better suited for binary variables. The performance of each test item was tested by item-total correlations (RIT) and item-rest correlations (RIR). Here also a deflation-corrected item-total correlation (DIT) and a deflation-corrected item-rest correlation (DIR) were calculated as they are more suited for binary variables. All test items with a DIR lower than 0.3 were excluded from the rest of the analyses as underperforming test items.

Structural validity of the CT test was examined by performing a confirmatory factor analysis (CFA), where the model compatibility to the data was tested for 2 models: a 1-factor model (all test items reflect the same underlying construct of computational thinking) and a 2-factor model (test items separate to 2 factors, ALG items and OTH items). Cross-country structural validity of the CT test was examined by first deciding a reference country (Lithuania), data of which was used as a reference data. CFA was applied to this reference data, and the CFA model was adjusted if needed so that the model indicators were as good as possible. After this the same CFA model was fitted to the data of other countries to see if the structural validity applies across countries.

The hypothesis of an underlying construct of mathematical thinking behind both computational thinking (which CT test reflects) and algebraic thinking (which AT test reflects) was examined by first making a new data table which had answers to both CT test items and AT test items for all individuals who had taken both tests. After this the association between the tests was examined by calculating the correlation between the total scores of both tests. Then the structural validity of a 2-factor model where each item belonged either to a computational thinking factor or an algebraic thinking factor was tested by fitted the said CFA model to the data.

IRT (Item Response Theory) analysis was used to examine the difficulty level and the discrimination ability of the test items. Quantiles (deciles) and a boxplot of the test total scores were calculated both for all individuals and separately for each country, but these descriptive statistics are not included in this summary. DWLS estimator was used in all CFA models, as it is more suitable for binary variables.

Pilot 2 CT results

Pilot 2 CT reliability summary

Deflation-corrected Cronbach's alpha coefficients indicated good test reliability across all age groups (alpha > 0.8). Two test items in age group 1 and two test items in age group 2 were excluded from further analyses due to underperformance (deflation-corrected item–rest correlation [DIR] < 0.3). These test items are shown in Table 1.

Table 1. Pilot 2 CT test items excluded from further analyses after reliability analysis.

Age Group 1	ALG.13.B (DIR = 0.29)	OTH.09.A (DIR = 0.28)
Age Group 2	ALG.03.A (DIR = 0.26)	OTH.08 (DIR = 0.11)
Age Group 3	-	-

Pilot 2 CT CFA summary

Confirmatory factor analysis suggested that for all age groups the test items do not reflect 2 underlying concepts of algorithmic skills (ALG) and algorithmic and other skills (OTH), but that all test items reflect one underlying concept of computational thinking. CFA also suggested that the test works better with older age groups, as they had better model diagnostic figures as age group 1. The set criteria (except chi-square test p-value) for good structural validity were met in age groups 2 and 3 for the 1 factor model, for the 1-factor model in age group 1 the set criteria were not met, but the model indicators were on a level that can be seen as acceptable (CFI approximately 0.9, RMSEA 0.06).

Pilot 2 CT cross-country validity summary

Due to limited country- and age-specific sample sizes, only country–age group combinations with sample sizes near 100 were included in the analysis. For the CT age group 2 tests, the only tested country (Spain) met the criteria for good structural validity, suggesting good cross-country validity in these tests. However, only one country was tested. In the CT age group 3 test, 2 of 4 tested countries (Finland and Spain) showed good or acceptable validity, indicating only partial cross-country validity. These results should be interpreted with caution due to (1) limited country coverage and (2) small sample sizes.

Pilot 2 CT IRT summary

The test appeared to be relatively difficult for all age groups, with age group 1 finding it the most challenging and age group 3 the least. The test seemed to perform best for the oldest age group, as the difficulty parameter estimates were most evenly distributed within the range [-2, 2] for that group. Overall, the test items performed well across all age groups, with only one item in age group 3 showing a parameter estimate that could be considered abnormal.

Pilot 2 CT and AT interrelation summary

There was a moderate positive correlation between the CT test total score and the AT test total score in all age groups. Data also supported the correlation being higher in age group 3 than in age group 1. CFA suggested that age group 3 test works the best and the age group 1 test the least well (structural validity is strongest in the age group 3 test and weakest in the age group 1 test). The 7-factor model (factors being computational thinking and 6 subsections of algebraic thinking) had best model indicators in all age groups. The set criteria for a good structural validity were not met in any age group, in the age group 3 and to a lesser degree in the age groups 2 and 1 the model indicators indicated structural validity that can be considered acceptable. Correlation between CT and AT factors estimated by 2-factor CFA models suggested moderate or strong positive correlation in all age groups.

Pilot 2 CT answer times

Answer times for the CT test were also analysed. In this summary only the minimum, maximum, mean and standard deviation of test answer times in minutes for the whole CT test and 2 test subgroups for all participants are presented. These results are shown in Table 2. For these statistics some students with outlier answer times were removed, see the main Pilot 2 report for details.

Table 2. Answer time (in minutes) statistics for Pilot 2 CT test.

	Age Groups 1-3		
(min)	Min	Mean (SD)	Max
Algorithmic thinking skills	0.36	8.92 (3.30)	22.68
Algorithmic thinking & other	0.35	8.53 (3.34)	26.78
Whole CT test	1.02	17.44 (5.83)	44.82

Pilot 2: Algebraic Thinking (AT) test

Pilot 2 AT test data

Pilot 2 AT test data also came from six countries, covered three age groups, and had only one version per test item. Unlike the CT test with two subgroups, AT had six: generalized arithmetic; equivalence, equations, and inequalities; functional thinking; representation; transformation; and transversal skills. Age group 1 test had 54 test items, age group 2 test had 59 test items and age group 3 test had 63 test items. After processing, the final dataset included 157209 non-NA answers from 2664 students. See the Pilot 2 main report for data processing details.

Pilot 2 AT methods

Methods used in Pilot 2 AT test were essentially the same as those used for the Pilot 2 CT test. While 2-factor CFA models were used in the CT analysis (ALG and OTH items), 6-factor CFA models were used in the AT analysis (6 AT subgroups).

Pilot 2 AT results

Pilot 2 AT reliability summary

Both raw Cronbach’s alpha coefficients and deflation-corrected Cronbach’s alpha coefficients indicated good test reliability across all age groups (alpha > 0.8). Three test items in age group 1, five test items in age group 2 and four test items in age group 3 were excluded from further analyses due to underperformance (deflation-corrected item–rest correlation [DIR] < 0.3). These test items are shown in Table 3.

Table 3. Pilot 2 AT test items excluded from further analyses after reliability analysis.

Age group 1	AT123_2.1.2, DIR = 0.07	AT1_4.2.1 DIR = -0.11	AT1_4.1.1 DIR = 0.28		
Age group 2	AT123_2.1.2 DIR = 0.05	AT2_1.1.11 DIR = 0.21	AT2_4.1.3 DIR = 0.18	AT2_4.2.1 DIR = -0.04	AT2_4.2.4 DIR = 0.28
Age group 3	AT123_2.1.2 DIR = 0.15	AT3_3.1.1 DIR = 0.29	AT3_3.2.3 DIR = 0.25	AT3_6.4.1 DIR = 0.29	

Pilot 2 AT CFA summary

For all age groups the 6-factor model had better model indicators than the 1-factor model. CFA suggests that the test works better with older age groups, as they had better model diagnostic figures as age group 1. All of the set criteria for good structural validity were not met in any of the age groups for the 6-factor model, but the model indicators were on a level that can be seen as acceptable especially in age group 3, and to a lesser degree in age groups 2 and 1.

Pilot 2 AT cross-country validity summary

Due to limited country- and age-specific sample sizes, a reduced CFA model with 11 items and 3 factors was used across all age groups. Only country–age group combinations with sample sizes near 100 were included in the analysis. For the AT age group 2 test, the only tested country (Spain) met the criteria for good structural validity, suggesting good cross-country validity in the test. However, only one country was tested. In the AT age group 3 test, 3 of 4 countries (Sweden, Hungary and Spain) showed good or acceptable validity, indicating relatively good cross-country validity (Finland's estimated model had negative item variances which made the model unfeasible). These results should be interpreted with caution due to (1) limited country coverage, (2) small sample sizes, and (3) the use of a highly reduced model.

Pilot 2 AT IRT summary

The results are very similar to those of the Pilot 2 CT test. The test seems to be on the difficult side for all age groups, most to age group 1 and least to age group 3. Test seems to work best for the oldest age group, as the difficulty parameter estimates are most evenly distributed to range [-2, 2] in that age group. In all age groups the models seem to work well, because there were very few abnormal estimate values. The test items seemed to perform quite well in all age groups, because there were very few test items with discrimination parameter values under 0.65.

Pilot 2 AT answer times

Answer times for the AT test were also analyzed. In this summary only the minimum, maximum, mean and standard deviation of test answer times in minutes for the whole AT test and 6 test subgroups for all participants are presented. These results are shown in Table 4.

Table 4. Answer time (in minutes) statistics for Pilot 2 AT test.

	Age Groups 1-3		
(min)	Min	Mean (SD)	Max
1. Generalized arithmetic	0.00	5.57 (2.43)	31.87
2. Equations and inequalities	0.00	5.43 (2.77)	22.75
3. Functional thinking	0.00	3.56 (1.91)	12.77
4. Representation	0.00	1.91 (1.45)	10.11
5. Transformation	0.00	2.78 (1.68)	16.94
6. Transversal skills for AT	0.00	3.16 (2.04)	43.15
Whole AT test	0.7	22.41 (7.99)	65.8

Co-funded by the European Union

Pilot 2 Student surveys

The Pilot 2 CT test included a pre-survey on student background and a post-survey on perceived math skills and test experience. The AT test used the same surveys, but the pre-survey was done only by Sweden, Türkiye, and Finland's age group 2.

A total of 2933 students completed the CT pre-survey, and 2812 completed the post-survey. This summary presents only the means and standard deviations of CT post-survey responses across three age groups by year, shown in Table 5. Survey questions were in a 5-point Likert scale where 1 meant "strongly disagree" and 5 meant "strongly agree".

Table 5. Means and standard deviations for Pilot 2 CT survey items for 3 groupings by age in years.

	Under 11 years		11-12 years		Over 12 years	
	Mean	Sd	Mean	Sd	Mean	Sd
Q1. I am good at math.	3.79	1.06	3.50	1.09	3.29	1.17
Q2. I did my best to complete the test.	4.28	1.04	4.20	0.99	3.97	1.10
Q3. The test was easy for me.	3.40	1.08	3.34	0.99	3.21	1.08
Q4. It was easy for me to complete the test in this digital system.	3.55	1.14	3.68	1.14	3.65	1.18
Q5. The test helped me to evaluate my skills.	3.80	1.25	3.49	1.24	3.06	1.26
Q6. I would like to take a similar test in the future.	3.66	1.38	3.21	1.44	2.87	1.50

A total of 425 students completed the AT pre-survey, and 2024 completed the post-survey. Survey questions were in the same 5-point Likert scale as CT survey questions. This summary presents only the means and standard deviations of AT post-survey responses across three test age groups, shown in Table 6.

Table 6. Means and standard deviations for Pilot 2 AT survey items by test age groups.

	Age group 1		Age group 2		Age group 3		All students	
	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
Q1. I am good at math.	3.63	1.14	3.36	1.14	3.17	1.22	3.37	1.18
Q2. I did my best to complete the test.	4.18	1.12	4.03	1.07	3.68	1.16	3.95	1.14
Q3. The test was easy for me.	3.04	1.21	3.04	1.12	2.93	1.23	3.00	1.19
Q4. It was easy for me to complete the test in this digital system.	3.44	1.24	3.43	1.21	3.43	1.29	3.43	1.25
Q5. The test helped me to evaluate my skills.	3.75	1.29	3.26	1.30	2.97	1.34	3.31	1.35
Q6. I would like to take a similar test in the future.	3.32	1.50	2.90	1.50	2.64	1.52	2.94	1.53

Co-funded by
the European Union

Pilot 2 Teacher survey

Teachers involved in Pilot 2 answered a survey covering their background and views on the CT and AT tests. The survey was completed by 45 teachers. In this summary only means and standard deviations of teachers' answers considering the CT test (Table 7) and AT test (Table 8) are presented. Survey questions were in a 5-point Likert scale where 1 meant "strongly disagree" and 5 meant "strongly agree".

Table 7. Means and standard deviations for Pilot 2 CT teachers survey items.

Survey question	Mean	Sd
Q1: The test measured students' computational thinking skills well.	4.23	0.75
Q2: The content of the test is in accordance with our national curriculum.	3.95	0.84
Q3: The difficulty level of the test was appropriate for most of my students.	3.98	0.86
Q4: The test does not need to be modified.	3.81	1.18
Q5: I found it easy to set up the test in this digital system.	4.52	0.71
Q6: The test helped me to evaluate my students' abilities.	4.14	0.89
Q7: It was nice to organize the test for my students.	4.51	0.70
Q8: I would like to organize similar tests for my students in the future.	4.56	0.80

Table 8. Means and standard deviations for Pilot 2 AT teachers survey items.

Survey question	Mean	Sd
Q1: The test measured students' algebraic skills well.	4.02	0.89
Q2: The content of the test is in accordance with our national curriculum.	3.78	0.82
Q3: The difficulty level of the test was appropriate for most of my students.	3.64	0.98
Q4: The test does not need to be modified.	3.56	1.27
Q5: I found it easy to set up the test in this digital system.	4.33	0.95
Q6: The test helped me to evaluate my students' abilities.	3.93	0.94
Q7: It was nice to organize the test for my students.	4.49	0.84
Q8: I would like to organize similar tests for my students in the future.	4.42	0.97