



# Analyzing–Evaluating–Creating: Assessing Computational Thinking and Problem Solving in Visual Programming Domains

Ahana Ghosh  
MPI-SWS  
Saarbrücken, Germany  
gahana@mpi-sws.org

Liina Malva  
MPI-SWS  
Saarbrücken, Germany  
Tallinn University  
Tallinn, Estonia  
liina.malva@tlu.ee

Adish Singla  
MPI-SWS  
Saarbrücken, Germany  
adishs@mpi-sws.org

## ABSTRACT

Computational thinking (CT) and problem-solving skills are increasingly integrated into K-8 school curricula worldwide. Consequently, there is a growing need to develop reliable assessments for measuring students’ proficiency in these skills. Recent works have proposed tests for assessing these skills across various CT concepts and practices, in particular, based on multi-choice items enabling psychometric validation and usage in large-scale studies. Despite their practical relevance, these tests are limited in how they measure students’ computational creativity, a crucial ability when applying CT and problem solving in real-world settings. In our work, we have developed ACE, a novel test focusing on the three higher cognitive levels in Bloom’s Taxonomy, i.e., ANALYZING, EVALUATING, and CREATING. ACE comprises a diverse set of 7×3 multi-choice items spanning these three levels, grounded in elementary block-based visual programming. We evaluate the psychometric properties of ACE through a study conducted with 371 students in grades 3–7 from 10 schools. Based on several psychometric analysis frameworks, our results confirm the reliability and validity of ACE. Our study also shows a positive correlation between students’ performance on ACE and performance on *Hour of Code: Maze Challenge* by Code.org.

## CCS CONCEPTS

• **Social and professional topics** → **Student assessment.**

## KEYWORDS

Computational Thinking, Assessment Tools, Bloom’s Taxonomy

### ACM Reference Format:

Ahana Ghosh, Liina Malva, and Adish Singla. 2024. Analyzing–Evaluating–Creating: Assessing Computational Thinking and Problem Solving in Visual Programming Domains. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE 2024)*, March 20–23, 2024, Portland, OR, USA. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3626252.3630778>

## 1 INTRODUCTION

Computational thinking (CT) is emerging as a critical skill in today’s digital world. According to the work of [30], “computational

thinking involves solving problems, designing systems, and understanding human behavior, by drawing on the concepts fundamental to computer science”. Several works have also discussed the multi-faceted nature of CT and its broader role in the acquisition of creative problem-solving skills [4, 10]. As a result, CT is being increasingly integrated into K-8 curricula worldwide [11, 16]. With the growing integration of CT at all academic stages, there has also been a surge in demand for validated and reliable tools to assess CT skills, especially at the K-8 stages [13, 15]. These assessment tools are essential for tracking the progress of students, guiding the design of curricula, and supporting teachers as well as researchers to assist students in the acquisition of CT skills [4, 9, 15, 31].

Prior work has proposed several assessments that measure students’ CT during their K-8 academic journey. On the one end, several portfolio-based assessments have been proposed that measure students’ CT through projects in specific programming environments [26]. Although portfolio-based tests provide *open-ended projects* to capture students’ analytical, evaluative, and creative skills, they are challenging to implement and interpret on a larger scale [13, 22]. On the other end, several diagnostic assessment tools have been proposed that measure CT in the form of *multiple-choice items* [2, 13, 14, 23]. These assessment tools are preferred for their practicality in large-scale administration and suitability for both pretest and posttest conditions [22]. However, scalability comes at the cost of limiting the ability to effectively measure students’ computational creativity. Thus, there is a need to develop multi-choice tests that also capture students’ computational creativity.

To this end, we have developed a novel test for grades 3–7, ACE, that focuses on the three higher cognitive levels of Bloom’s Taxonomy, i.e., ANALYZING, EVALUATING, and CREATING [1]. It comprises a diverse set of multiple-choice items spanning all three higher cognitive levels, including the highest level of CREATING. Figure 1 illustrates the diversity of items covered by ACE. Further details of the development of ACE are presented in Section 3. In this paper, our objective is to validate ACE with students from grades 3–7, and report on its psychometric properties. Specifically, we center the analysis around the following research questions: (1) **RQ1**: How is the internal structure of ACE organized w.r.t. item categories pertaining to Bloom’s higher cognitive levels? (2) **RQ2**: What is the reliability of ACE w.r.t. consistency of its items? (3) **RQ3**: How does performance on ACE correlate with performance on real-world programming platforms and students’ prior programming experience?




This work is licensed under a Creative Commons Attribution International 4.0 License.

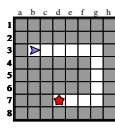
SIGCSE 2024, March 20–23, 2024, Portland, OR, USA  
© 2024 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-0423-9/24/03.  
<https://doi.org/10.1145/3626252.3630778>

Level \ Concept	APPLYING-ANALYZING	ANALYZING-EVALUATING	EVALUATING-CREATING
	Solution checking Solution tracing	Code debugging Code equivalence	AVATAR GOAL design WALL design
Basic moves and turns	Q01	Q08	Q15
REPEAT{}	Q02 Q05	Q12	Q16
REPEATUNTIL{}	Q06	Q09	Q17 Q18
REPEATUNTIL{IF}	Q07	Q10	Q19
REPEATUNTIL{IFELSE}	Q04	Q11 Q14	Q20, Q21
REPEAT{REPEAT}		Q13	
REPEAT{IF}	Q03		

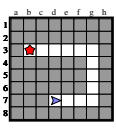
(a) Our Test: Distribution of Items

Q07. You are given a code. You are also given three grids GRID-1, GRID-2, and GRID-3. Which of these grids can be solved with this code?

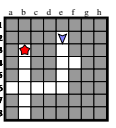




GRID-1



GRID-2

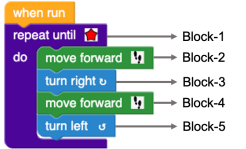


GRID-3

OPTION A	Only GRID-1
OPTION B	GRID-1 and GRID-2
OPTION C	GRID-1 and GRID-3
OPTION D	All three grids GRID-1, GRID-2, and GRID-3

(b) Q07. Solution checking

Q09. You are given a code and a grid. You may have to fix some errors in the code such that it solves the grid. How can you fix the code?



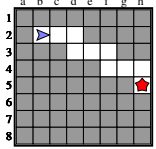
Block-1 →

Block-2 →

Block-3 →

Block-4 →

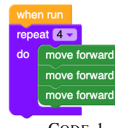
Block-5 →



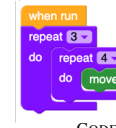
OPTION A	The code does not have any errors and it already solves the grid
OPTION B	Add <code>move forward</code> after Block-2
OPTION C	Add <code>move forward</code> after Block-4
OPTION D	Change Block-3 to <code>turn left</code> and Block-5 to <code>turn right</code>

(c) Q09. Code debugging

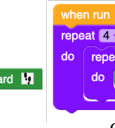
Q13. You are given a code CODE-1 and two smaller codes CODE-2 and CODE-3. You have to think about the AVATAR's behavior when a code is run on a grid. Which of these two smaller codes produce the same behavior as CODE-1 on any grid?



CODE-1



CODE-2




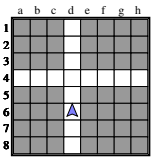
CODE-3

OPTION A	None of these two smaller codes
OPTION B	Only CODE-2
OPTION C	Only CODE-3
OPTION D	Both CODE-2 and CODE-3

(d) Q13. Code equivalence

Q18. You are given a code and an incomplete grid without the GOAL. You can add the GOAL in any grid cell which is not occupied by the AVATAR and is not a WALL. How many different locations of the GOAL are possible such that the grid is solved by the code?




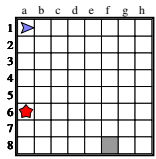


OPTION A	1
OPTION B	4
OPTION C	5
OPTION D	8

(e) Q18. GOAL design

Q21. You are given a code and an incomplete grid. You can add additional WALL cells to the grid by converting any of FREE cells into WALL cells. What is the smallest number of additional WALL cells you must add such that the grid is solved by the code?





OPTION A	1
OPTION B	2
OPTION C	7
OPTION D	8

(f) Q21. WALL design

Figure 1: (a) shows the distribution of test items w.r.t to CT and problem-solving concepts and Bloom's cognitive levels. (b)–(f) are examples of five items from ACE. These items are grounded in the domain of *Hour of Code: Maze Challenge* (HoCMAZE) [6], which can be found at [studio.code.org/s/hourofcode](https://studio.code.org/s/hourofcode). HoCMAZE domain comprises elementary block-based visual programming tasks where one has to write a solution code that would navigate the AVATAR (blue dart) to the GOAL (red star) without crashing into WALLS (gray grid cells). We encourage the reader to attempt these items; answers are at the end of page 6.

## 2 RELATED WORK

Prior work has proposed several CT assessments and their categorizations based on their format including the following [13, 22]: (a) *portfolios*, which are project-based programming assessments; (b) *interviews*, which are used in conjunction with portfolios to

gain insights into students' thinking process; (c) *summative assessments*, which are long-format answer type questions to measure CT specifically in the context of a particular domain; (d) *multi-choice diagnostic tests*, which measure CT aptitude and may be administered in both pretest and posttest conditions. As mentioned in Section 1,

**Table 1: Categorization of different CT assessments proposed in recent works. The first column shows the specific CT Assessment. The next three columns, APPLYING-ANALYZING, ANALYZING-EVALUATING, and EVALUATING-CREATING, classify the assessment based on these different cognitive levels of Bloom’s Taxonomy where “✓” implies presence of the levels and “✗” implies absence of the levels. The “Grade” column refers to the intended grades (age group) for the test. The “Validity” column refers to three dimensions across which the test was validated, including (i) “Student”: test items validated with students; (ii) “Expert”: test items validated with experts; (iii) “Convergent”: test validated w.r.t. performance on another test/course. Finally, the “Domain” column shows the domain on which the items in the test were designed. Further details are presented in Section 2.**

CT Assessment and Tests	APPLYING-ANALYZING	ANALYZING-EVALUATING	EVALUATING-CREATING	Grades	Validity:			Domain
					Student	Expert	Convergent	
ACE (this paper)	✓	✓	✓	3-7	✓	✓	✓	block-based visual programming
cCTt [13]	✓	✓	✗	3-4	✓	✓	✗	block-based visual programming
TechCheck [21]	✓	✓	✗	K-4	✓	✓	✓	everyday scenarios
CT-Test [9, 22, 23]	✓	✓	✗	6-8	✓	✓	✓	block-based visual programming
Gane et al. 2021 [8]	✓	✓	✗	3-4	✓	✓	✗	block-based visual programming and everyday scenarios
Chen et al. 2017 [3]	✓	✓	✗	5	✓	✓	✗	robotics programming and everyday scenarios
ACES [18]	✓	✓	✗	3-5	✓	✗	✗	block-based visual programming and everyday scenarios
CTC [14]	✓	✓	✗	8-12	✓	✓	✓	block-based visual programming and real-world problem-solving
Commutative Assessment [28]	✓	✓	✗	8-12	✓	✓	✗	block-based visual programming and text-based visual programming
Mühling et al. 2015 [17]	✓	✓	✗	8-10	✓	✗	✗	similar to Karel programming [19]
PSIv1 [2]	✓	✗	✗	13-14 (college)	✓	✓	✓	text-based programming

we focus on multi-choice CT tests due to their practicality and scalability. Table 1 presents several different multi-choice diagnostic tests proposed in the literature, viewed through the lens of Bloom’s taxonomy [22]. Specifically, we classify them based on their coverage of the higher cognitive levels of the taxonomy (APPLYING, ANALYZING, EVALUATING, and CREATING).

These tests cater to students from different school years, starting from kindergarten (K) through the early years of college. Next, we describe three representative assessments in different years. The *competent Computational Thinking test* (cCTt) [13] was proposed recently in 2022 for students in grades 3–4. The test comprised items that only required finding solution codes or completing a given solution code. These types of items invoke students’ APPLYING, ANALYZING, and EVALUATING cognitive levels. The *Computational Thinking Challenge* (CTC) [14] was proposed in 2021 for students in grades 9–12. The test contains programming items in the form of Parsons problems [32], solution-finding multi-choice items, and general items on real-world problem-solving. The items in CTC also cover all cognitive levels except the CREATING level. Finally, *Placement Skill Inventory v1* (PSIv1) [2] was also proposed recently in 2022 for college students as a placement test. The test contains multi-choice theoretical items on programming and covers only the APPLYING and ANALYZING levels of Bloom’s taxonomy. Contrary to these tests, ACE contains items that require synthesizing new problem instances to verify the correctness of a proposed solution. These items in ACE are intended to cover the Bloom’s CREATING cognitive level. ACE is developed for students in grades 3–7.

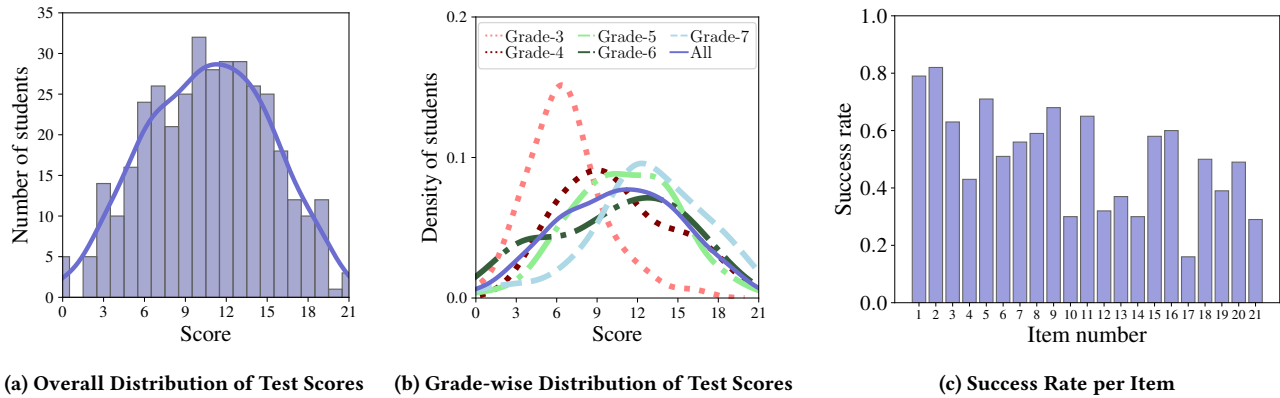
Table 1 also shows different domains in which CT is measured in these tests. For grades K–8, the most popular setting is block-based visual programming, likely because of the low syntax overhead of the domains and the ease of measuring CT concepts such as

conditionals, loops, and sequences [9, 13, 23, 28]. Beyond block-based programming domains, several CT tests also utilize real-world settings, including everyday-scenarios (e.g., a scenario related to seating arrangements in a gathering) [18], robotics [3], and real-world problem-solving (e.g., a problem related to route planning in a city) [14]. The advantage of these real-world settings and domains is their administration with minimal domain knowledge, thus making them suitable pretest and posttest candidates. ACE is based on the block-based visual programming domain.

Finally, an important aspect of developing such CT assessments is their validation and reliability [13]. Generally, CT assessments are validated using three methods: (a) with students in specific grades for which the assessment was designed; (b) with expert feedback; (c) w.r.t. another test or performance in a course (i.e., convergent validity). For a well-rounded evaluation, it is advisable to explore all three validation methods [13, 22]. As shown in Table 1, most tests are validated with students, while some of them are refined by experts. However, the incorporation of convergent validity is less common. ACE is validated using all three methods.

### 3 OUR TEST: ACE

The development of ACE is centered around the higher cognitive levels of Bloom’s taxonomy: ANALYZING, EVALUATING, and CREATING. The test contains items grounded in the domain of block-based visual programming. Specifically, we consider the popular block-based visual programming domain of *Hour of Code: Maze Challenge* [6] by code.org [5]. We picked this domain as it encapsulates important CT and problem-solving concepts of conditionals, loops, and sequences, within the simplicity of the block-based structure. Students can attempt tasks in this domain with a simple description of the constructs and task, as discussed in the caption of Figure 1.



**Figure 2: An overview of the performance of students on ACE. (a) overall distribution of ACE scores across all 371 students; (b) distribution of ACE score per grade; (c) success rate of students for each item in ACE. Details are in Section 4.**

Next, we describe the items in ACE which are divided into the following three categories based on Bloom’s higher cognitive levels:

- **APPLYING–ANALYZING:** This category comprises items either on finding a solution code of a given task or reasoning about the trace of a given solution code on one or more visual grids. They are based on the **APPLYING** and **ANALYZING** levels of Bloom’s taxonomy, as they require applying CT concepts and analyzing code traces. These items are typically the most common type of items included in several CT tests [13, 18].
- **ANALYZING–EVALUATING:** This category comprises items that require reasoning about errors in candidate solution codes of a task and evaluating the equivalence of different codes for a given task. They are based on the **ANALYZING** and **EVALUATING** levels of Bloom’s taxonomy. Several CT assessments also include these types of debugging items [9, 14].
- **EVALUATING–CREATING:** This category comprises items that require reasoning about the design of task grids for given solution codes. They are based on the **EVALUATING** and **CREATING** levels of Bloom’s taxonomy, as they involve synthesizing components of visual grids such as **AVATAR**, **GOAL**, and **WALL**. These items are unique to ACE and capture the open-ended nature of task design, such as counting several possible task configurations to satisfy a given solution code (see items Q18 and Q21 in Figure 1).

In the process of developing ACE, we consulted with CS educators and researchers with expertise in using CT tools for K-12 education. Five experts provided feedback on an initial version of the test in terms of items’ suitability and difficulty. Furthermore, we asked six students (not part of the study) from grades 3 to 6 to attempt a version of the test while recording their thought processes via think-aloud methods. The responses of these students allowed us to further refine the phrasing and structure of the items. Ultimately, the final version of ACE contained 21 single-correct multiple-choice items to be completed in one 45-minute lesson. The items were divided equally across three categories based on Bloom’s cognitive levels, i.e., each category comprised 7 items. The three categories are henceforth referred to as ACE[01-07], ACE[08-14], and ACE[15-21], respectively. Figure 1 presents the breakdown of all our 21 test items and also illustrates 5 items from ACE.

## 4 STUDY AND DESCRIPTIVE STATISTICS

In this section, we provide details of the data collection process for ACE’s psychometric evaluation.

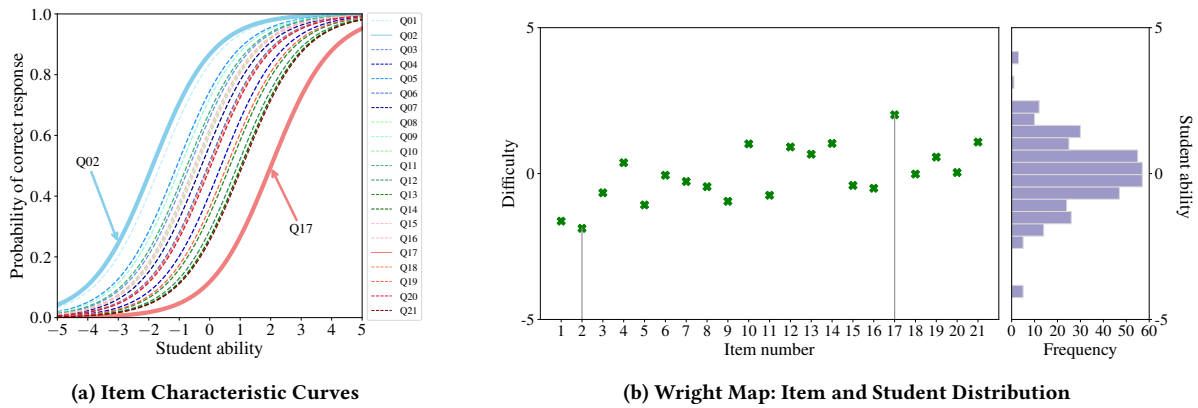
### 4.1 Two-Phase Data Collection Process

The study to evaluate the psychometric properties of ACE was planned in two phases, spread across two weeks. The first phase was intended to familiarize students with the block-based visual programming domain of *Hour of Code: Maze Challenge* (HoCMaze) [6] by code.org [5], and introduce them to basic programming concepts. Additionally, it would serve as a baseline to correlate students’ performance w.r.t. ACE, and measure the convergent validity of ACE. In the second phase of the study, students would take the ACE test. This two-phase study design ensured that students would have enough focus on each study component as well as a time gap between domain familiarity and the actual test.

We obtained an Ethical Review Board approval from the Ethics Committee of Tallinn University before conducting the study. The study was conducted in Estonia, where a random selection of 10 schools was pooled from 11 out of 15 counties. Participation in the study was voluntary for both teachers and students. Next, we outline the details of each phase of the study.

The data collection process was conducted in May 2023. During both phases, students received usernames to ensure anonymity throughout the study. The first phase of data collection included one 45-minute lesson during which the students filled in a short background questionnaire in Google Forms (about 5 minutes) and then solved 20 tasks from HoCMaze (about 40 minutes). We hosted these 20 tasks on a separate platform created for the study to enable the collection of students’ performance data on these tasks. Students were allowed multiple attempts to solve each task and could score a maximum of 20 points, i.e., 1 point per task. Henceforth, we refer to students’ performance in this phase as their HoCMaze score. The second phase took place one week later and involved a 45-minute lesson during which students took the ACE test. The test was administered through a Qualtrics survey. Students could score a maximum of 21 points, i.e., 1 point per item. Henceforth, we refer to students’ performance in this phase as their ACE score.





**Figure 3: Results from a 1-parameter Rasch model [20] on the ACE items and student scores. (a) Item characteristic curve for each item in ACE and (b) Wright map corresponding to our student population.**

## 4.2 Participant Details and Descriptive Statistics

A total of 371 students participated in the study, distributed across five different grades as follows:  $n = 51$  from grade 3,  $n = 34$  from grade 4,  $n = 114$  from grade 5,  $n = 81$  from grade 6, and  $n = 91$  from grade 7. The participants' age varied from 9 to 15 years. The distribution of students w.r.t. gender was as follows: 48% girls and 52% boys. Regarding the students' prior programming experience, their reported duration of programming experience was as follows: 40% reported no programming experience, 22% reported at least 1 year, 15% reported at least 2 years, 13% reported at least 3 years, and 10% reported between 4 to 6 years. Among the 60% of participants ( $n = 223$ ) having programming experience, they reported the following sources of gaining programming experience (multiple sources could be selected): 77% studied programming in school lessons, 31% studied programming in after-school lessons, and 14% studied programming independently at home.

Figure 2 summarizes the students' performance on ACE. The average score of all 371 students is 10.69. The distribution of the scores across all participants is shown in Figure 2a. Grade 7 averaged the highest score of 12.84, while grade 3 averaged the lowest score of 6.67. Grades 4, 5, and 6 had similar average scores of around 10.5. Figure 2b shows the distribution of scores per grade, and Figure 2c shows the success rate per item. Overall, the success rate of ACE[01-07] was higher than those of ACE[08-14] and ACE[15-21]. More concretely, the average score of all 371 students when measured per category (i.e., points on 7 items in each category) was as follows: 4.45 on ACE[01-07], 3.22 on ACE[08-14], and 3.02 on ACE[15-21].

## 5 RESULTS AND DISCUSSION

In this section, we discuss the results of the study centered around the research questions (RQs) introduced in Section 1.

### 5.1 RQ1: Internal Structure of ACE

We assess the internal structure of ACE w.r.t. its three item categories (ACE[01-07], ACE[08-14], ACE[15-21]) as its underlying factors using Confirmatory Factor Analysis (CFA), a standard method in quantitative test analysis [7, 14]. CFA determines whether the structure of ACE scores aligns with the three item categories as

three factors. Specifically, CFA provides the Root Mean Square Error of Approximation (RMSEA) as the goodness of fit for statistical models. RMSEA values between 0 and 0.01 indicate excellent fit, and values up to 0.05 indicate good fit. Additionally, CFA provides the Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI). CFI measures how well our three-factor model fits the observed data on test scores compared to an independent-item model with 21 items as factors. TLI also measures model fit by accounting for model simplicity. CFI and TLI values greater than 0.90 indicate a good fit.

Our test presented a significant model with good fit statistics ( $p < 0.01$ ; RMSEA=0.0275; CFI=0.945; TLI=0.938). This analysis also highlighted a potential issue with item Q17, as the Standardized Estimate for CFA's factor loading of this item seemed problematic with a value of  $-0.111$  ( $p = 0.059$ ). Nevertheless, the model fit did not improve significantly after removing item Q17 from the data; new statistics without Q17 ( $p < 0.01$ ; RMSEA=0.0278; CFI=0.949; TLI=0.942) are similar to statistics reported above without any significant difference. We discuss this item further as part of RQ2.

### 5.2 RQ2: Reliability of ACE

Next, we determine the reliability of ACE, i.e., a measure of its ability to produce consistent and stable results over repeated administrations (a higher value being better). One standard way to measure this is through the Cronbach alpha value [14, 27] that reflects the average inter-item correlations in a test. Another method is the reliability of student ability estimates obtained from Item Response Theory (IRT). In our study, we apply IRT analysis on students' responses to ACE and fit a 1-parameter logic Rasch model (1-PL IRT) [20]. The model estimates the per-item difficulties and students' abilities, and provides the reliability of these estimates.

The overall reliability for our test was good with a Cronbach alpha value of 0.813. Among the three item categories, Cronbach alpha was 0.622 for ACE[01-07], 0.562 for ACE[08-14], and 0.625 for ACE[15-21]. Figure 3a shows the 1-PL IRT item characteristic curves for all items; we find that Q02 is the easiest and Q17 is the hardest ACE item. Figure 3b illustrates the difficulty of items as well as the estimated ability of students' in our population. The 1-PL IRT Person reliability value for all 21 items is 0.790 (with  $p < 0.01$ ).

Q17. You are given a code and an incomplete grid without the AVATAR. How many different positions of the AVATAR are possible such that the grid is solved by the code?

```

when run
repeat until
do
  move forward
  turn left
  move forward
  turn right
        
```

	a	b	c	d	e	f	g	h
1								
2								
3								
4								
5								
6								
7								
8								

OPTION A	1
OPTION B	2
OPTION C	4
OPTION D	More than 4

Figure 4: Q17. AVATAR design

Next, we discuss the potentially problematic item Q17 shown in Figure 4. We find that its exclusion from the model doesn't significantly improve the IRT Person reliability. One possible reason Q17 prompted incorrect responses is that it was the first item in ACE requiring enumeration of all possible AVATAR locations. However, students adapted to similar formats in subsequent items (e.g., Q18 and Q21 in Figure 1). Prior work confirms that varying response formats can cause deviations [12]. A possible revision of item Q17 could be simplifying the visual grid to reduce its complexity.

### 5.3 RQ3: Correlating ACE scores

We measure the convergent validity of ACE w.r.t. HoCMAZE scores. Additionally, we also measure the correlation of the three ACE categories with both HoCMAZE scores as well as overall ACE scores. Finally, we measure the influence of extrinsic factors such as prior programming experience on ACE scores. To measure all these correlations, we perform standard Pearson's correlation analysis between each of these features on data from our entire student population [14, 24]. High positive values of Pearson's correlation coefficient,  $r$ , indicate a strong positive correlation.

The correlation results for ACE with HoCMAZE and item categories of ACE are shown in Figure 5. For instance, the Pearson correlation between ACE scores and HoCMAZE scores w.r.t. all students is  $r = 0.41$ ,  $p < 0.001$ , confirming the convergent validity of ACE. All item categories were (significantly) positively correlated with ACE and also have high positive inter-category correlations.

In terms of the effect of prior programming experience on ACE, we observed a significant positive correlation with both the student's year of study ( $r = 0.358$ ,  $p < 0.01$ ) and age ( $r = 0.359$ ,  $p < 0.01$ ). Our result aligns with prior work [25] indicating that participants' developmental factors (e.g., reading skills, abstract thinking) can impact test performance. In our student population, varying programming exposure due to elective programming courses influenced prior programming skills. Analyzing this further, we discovered that students who took after-school programming classes outperformed those who did not on ACE ( $p < 0.05$ , w.r.t.  $t$ -test [29]).

### 5.4 Limitations

Next, we discuss a few limitations of our current study. Firstly, in this study, we evaluated the convergent validity of ACE w.r.t. HoCMAZE scores. However, it would be more informative to evaluate

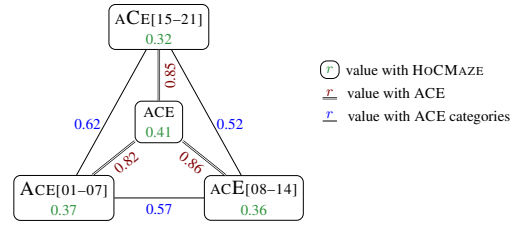


Figure 5: Pearson's correlation coefficient,  $r$ , between ACE and HoCMAZE, between ACE and its categories, and between each category. All values are significant with  $p < 0.001$ .

ACE w.r.t. other types of assessments, such as portfolios, which specifically consider CREATING cognitive level. Moreover, it would be interesting to evaluate the convergent validity of ACE w.r.t. students' performance in other subjects involving CT. Secondly, grade 3 did not present a significant correlation between ACE and HoCMAZE scores (Pearson's  $r = 0.068$ ;  $p = 0.633$ ), possibly because of difficulties with text comprehension of the item descriptions. Hence, refining the presentation of items could be beneficial for this age group. Finally, we presented the test items in a fixed order, which might have affected students' performance on specific items such as Q17. Implementing a randomized order of the test items within each category could be a way to address this limitation.

## 6 CONCLUSION AND FUTURE WORK

We developed a new test, ACE, to assess CT and problem-solving skills, focusing on higher levels of Bloom's taxonomy, including CREATING. We capture this level through a novel category of items that go beyond solution finding or debugging and consider task design. In this paper, we studied the psychometric properties of ACE, and our results confirm ACE's reliability and validity. There are several exciting directions for future work. Firstly, we can extend the framework of items to develop tests with more advanced programming constructs, such as variables/functions suitable for higher grades. Secondly, while we studied the utility of items in ACE for CT assessments, these items could also be incorporated as part of the curriculum to teach students richer CT and problem-solving skills such as problem design and test-case creation.

Answers to Q07, Q09, Q13, Q18, Q21 (Figure 1) and Q17 (Figure 4) are as follows: C, B, D, C, A, and D, respectively.

## ACKNOWLEDGMENTS

We would like to thank the teachers and students in Estonia for their participation in the study. We thank Danial Hoosyar and the Ethics Committee of Tallinn University, Estonia, for reviewing and providing the Ethical Review Board (ERB) approval for the study. Ahana Ghosh acknowledges support from Microsoft Research through its PhD Scholarship Programme. Funded/Cofunded by the European Union (ERC, TOPS, 101039090). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

## REFERENCES

- [1] Benjamin S Bloom and David R Krathwohl. 2020. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Book 1, Cognitive Domain*.
- [2] Ryan Bockmon and Chris Bourke. 2023. Validation of the Placement Skill Inventory: A CS0/CS1 Placement Exam. In *SIGCSE 2023*.
- [3] Guanhua Chen, Ji Shen, Lauren Barth-Cohen, Shiyang Jiang, Xiaoting Huang, and Moataz Eltoukhy. 2017. Assessing Elementary Students' Computational Thinking in Everyday Reasoning and Robotics Programming. *Computers & Education* (2017).
- [4] Morgane Chevalier, Christian Giang, Alberto Piatti, and Francesco Mondada. 2020. Fostering Computational Thinking Through Educational Robotics: A Model for Creative Computational Problem Solving. *International Journal of STEM Education* (2020).
- [5] Code.org. 2022. Code.org. <https://code.org/>.
- [6] Code.org. 2022. Hour of Code: Classic Maze Challenge. <https://studio.code.org/s/hourofcode>.
- [7] Christine DiStefano and Brian Hess. 2005. Using Confirmatory Factor Analysis for Construct Validation: An Empirical Review. *Journal of Psychoeducational Assessment* (2005).
- [8] Brian D. Gane, Maya Israel, Noor Elagha, Wei Yan, Feiya Luo, and James W. Pellegrino. 2021. Design and Validation of Learning Trajectory-Based Assessments for Computational Thinking in Upper Elementary Grades. *Computer Science Education* (2021).
- [9] Marcos Román González. 2015. Computational Thinking Test: Design Guidelines and Content Validation. In *EDULEARN 2015*.
- [10] Shuchi Grover, Satadri Basu, Marie A. Bienkowski, Michael Eagle, Nicholas Diana, and John C. Stamper. 2017. A Framework for Using Hypothesis-Driven Approaches to Support Data-Driven Learning Analytics in Measuring Computational Thinking in Block-Based Programming Environments. *Transactions on Computing Education* (2017).
- [11] Shuchi Grover and Roy Pea. 2013. Computational Thinking in K–12: A Review of the State of the Field. *Educational Researcher* 42, 1 (2013), 38–43.
- [12] Josef Guggemos, Sabine Seufert, and Marcos Román-González. 2023. Computational Thinking Assessment - Towards More Vivid Interpretations. *Technology, Knowledge and Learning* (2023).
- [13] Laila El Hamamsy, María Zapata-Cáceres, Estefanía Martín-Barroso, Francesco Mondada, Jessica Dehler-Zufferey, and Barbara Bruno. 2022. The competent Computational Thinking test (cCTt): Development and Validation of an Unplugged Computational Thinking Test for Upper Primary School. *CoRR* abs/2203.05980 (2022).
- [14] Rina P. Y. Lai. 2022. Beyond Programming: A Computer-Based Assessment of Computational Thinking Competency. *Transactions on Computing Education* (2022).
- [15] James Lockwood and Aidan Mooney. 2018. Computational Thinking in Secondary Education: Where Does it Fit? A Systematic Literary Review. *International Journal of Computer Science Education in Schools* (2018).
- [16] Dastyni Loksa, Amy J. Ko, Will Jernigan, Alannah Oleson, Christopher J. Mendez, and Margaret M. Burnett. 2016. Programming, Problem Solving, and Self-Awareness: Effects of Explicit Guidance. In *CHI*.
- [17] Andreas Mühling, Alexander Ruf, and Peter Hubwieser. 2015. Design and First Results of a Psychometric Test for Measuring Basic Programming Abilities. In *Workshop in Primary and Secondary Computing Education*.
- [18] Miranda C. Parker, Yvonne S. Kao, Dana Saito-Stehberger, Diana Franklin, Susan Krause, Debra J. Richardson, and Mark Warschauer. 2021. Development and Preliminary Validation of the Assessment of Computing for Elementary Students (ACES). In *SIGCSE*.
- [19] Richard E Pattis. 1981. *Karel the Robot: A Gentle Introduction to the Art of Programming*. John Wiley & Sons, Inc.
- [20] Georg Rasch. 1993. *Probabilistic Models for Some Intelligence and Attainment Tests*. Education Resources Information Center (ERIC).
- [21] Emily Relkin, Laura de Ruiter, and Marina Umaschi Bers. 2020. TechCheck: Development and Validation of an Unplugged Assessment of Computational Thinking in Early Childhood Education. *Journal of Science Education and Technology* (2020).
- [22] Marcos Román-González, Jesús Moreno-León, and Gregorio Robles. 2019. Combining Assessment Tools for a Comprehensive Evaluation of Computational Thinking Interventions. *Computational Thinking Education* (2019).
- [23] Marcos Román-González, Juan-Carlos Pérez-González, and Carmen Jiménez-Fernández. 2017. Which Cognitive Abilities Underlie Computational Thinking? Criterion Validity of the Computational Thinking Test. *Computers in Human Behavior* 72 (2017), 678–691.
- [24] Philip Sedgwick. 2012. Pearson's Correlation Coefficient. *British Medical Journal* (2012).
- [25] Alaaeddin Swidan, Feliene Hermans, and Marileen Smit. 2018. Programming Misconceptions for School Students. In *ICER*.
- [26] Xiaodan Tang, Yue Yin, Qiao Lin, Roxana Hadad, and Xiaoming Zhai. 2020. Assessing Computational Thinking: A Systematic Review of Empirical Studies. *Computers & Education* (2020).
- [27] Mohsen Tavakol and Reg Dennick. 2011. Making Sense of Cronbach's Alpha. *International Journal of Medical Education* (2011).
- [28] David Weintrop and Uri Wilensky. 2015. Using Commutative Assessments to Compare Conceptual Understanding in Blocks-based and Text-based Programs. In *ICER*.
- [29] Bernard L Welch. 1947. The Generalization of Student's Problem When Several Different Population Variances are Involved. *Biometrika* 34, 1-2 (1947), 28–35.
- [30] Jeannette M Wing. 2006. Computational Thinking. *Communications of the ACM* (2006).
- [31] María Zapata-Cáceres, Estefanía Martín-Barroso, and Marcos Román-González. 2020. Computational Thinking Test for Beginners: Design and Content Validation. In *EDUCON*.
- [32] Rui Zhi, Min Chi, Tiffany Barnes, and Thomas W. Price. 2019. Evaluating the Effectiveness of Parsons Problems for Block-based Programming. In *ICER*.