MEASURING COMPUTATIONAL THINKING – DEVELOPING A SHORT PERFORMANCE TEST FOR HIGHER EDUCATION

Josef Guggemos[1*], Roman Rietsche[2*], Stephan Aier[3], Jannis Strecker[3] and Simon Mayer[3]
[1]*Department of Vocational Education and Training, University of Education Schwäbisch Gmünd, Germany*
[2]*Institute for Digital Technology Management, Bern University of Applied Sciences, Switzerland*
[3]*School of Computer Science, University of St.Gallen, Switzerland*
**Josef Guggemos and Roman Rietsche contributed equally*

ABSTRACT

Technological advancements, particularly in artificial intelligence, significantly transform our society and work practices. Computational thinking (CT) has emerged as a crucial 21st-century skill, enabling individuals to solve problems more effectively through an automation-oriented perspective and fundamental concepts of computer science. To ensure the effective integration of CT into educational curricula, it is crucial to develop efficient assessment frameworks that allow teachers to measure and promote student CT proficiency. Therefore, our aim is to develop a short test to measure CT among undergraduate students. To this end, we consider two performance tests: the Computational Thinking test (CTt) and the Algorithmic Thinking Test for Adults (ATTA). We use items from both instruments to compile a short test. Based on a sample of 290 second-year non-computer science undergraduate students, we provide evidence on the quality of our test. Besides classical test theory, we apply item response theory, namely Rasch modeling, and confirmatory factor analysis. Our test shows favorable properties, e.g., Cronbach's alpha > .75, and may be suitable for the efficient assessment of CT across higher education programs.

KEYWORDS

Computational Thinking, Assessment, Rasch-scaling, Performance Test, Higher Education

1. INTRODUCTION

Far-reaching technological changes are shaping our society and ways of working (Zhang et al., 2024). Key drivers for these changes are advances in the field of computing (Wing, 2008). From an educational point of view, it is vital to determine and develop the skills necessary for success in such an environment. Among others, computational thinking (CT) is regarded as a key 21st-century skill (Voogt et al., 2015; Wing, 2006). Wing conceptualized CT as "solving problems, designing systems, and understanding human behavior, by drawing on the concepts fundamental to computer science" (2006, p. 33).

Much research has been carried out on CT in K–12 settings (Lu et al., 2022). This includes conceptualizations and delineations of CT (Shute et al., 2017), CT instruction (Hsu et al., 2018), and CT assessment (Tang et al., 2020). Although there is substantial evidence regarding CT in K–12 settings, research on CT in higher education is only gradually evolving (Zhang et al., 2024). Higher education should contribute to student employability (Cheng et al., 2022; Lu et al., 2022). Hence, in an increasingly digital environment, CT may be important to achieve this goal (ISTE, 2015).

A prerequisite for quantitative CT research (in higher education) are assessment instruments that can capture a student's level of CT and track its development (Lye & Koh, 2014). In comparison to self-assessments, that are subjective by nature, test instruments (tests) have the advantage that they objectively measure performance instead of self-efficacy (Lafuente Martínez et al., 2022). They offer an upper bound for students' actual proficiency (Bühner, 2011). CT tests are available for K–12 and adults (Lafuente Martínez et al., 2022; Román-González et al., 2017; Zhang et al., 2024); however, studies on the assessment of CT in higher education primarily address computer science students (Lu et al., 2022). Hence,

these tests may not be well-suited for students in other domains because they may be too difficult for such students.

Beyond its research applications, assessing student CT proficiency at the start of a course offers substantial benefits. It enables instruction tailored to individual learning needs, allowing lecturers to allocate their limited time more efficiently to students who require additional support. Moreover, tutors could focus on those students with lower test scores. Such measures can support study success (Ifenthaler & Yau, 2020). However, test time is often an issue in higher education (Schlax et al., 2020). Current CT tests typically require a minimum of 45 minutes to complete. Administering both a pre-test and a post-test would necessitate dedicating time equivalent to an entire lecture. This may be impractical within the constraints of a standard course schedule. Overall, there is a need for a short test to assess the CT proficiency of students in higher education outside the disciplines of STEM. Against this background, we raise the following research question:

RQ: How can the computational thinking proficiency of undergraduate students be tested efficiently?

To answer this question, section two reviews the conceptual basis of CT, the state of CT research in higher education, and suitable CT tests. We do not consider instruments for large-scale assessment, like the International Computer and Information Literacy Study (ICILS) (Fraillon et al., 2019). Instead, we focus on freely available tests that researchers and educators may use to measure CT in higher education. Due to the importance of CT interventional studies in higher education (Jong & Jeuring, 2020), instruments should be suitable for pre-post, or longitudinal designs. Section three presents our method and sample. Section four discusses the results.

Our research makes three substantial contributions. First, we offer a validated short test that efficiently assesses CT. This means it achieves an internal consistency reliability greater than .75 with only 11 dichotomous items. Second, we provide empirical evidence that our test items are suitable for undergraduate students in the social sciences, i.e., non-STEM. Third, our test ensures fairness and non-discrimination in CT assessment in terms of gender, which is essential for equitable education and research.

2. STATE OF THE ART

2.1 Computational Thinking

A core set of CT facets can be identified that specifies this concept (Lyon & Magana, 2020; Shute et al., 2017). Table 1 summarizes this core set.

Table 1. Core facets of CT

Abstraction	"simplifying from the concrete to the general as solutions are developed" (Barr & Stephenson, 2011, p. 52)
Algorithmic thinking	using "a step-by-step procedure for taking input and producing some desired output" (Wing, 2008, p. 3718)
Automation	"process in which a computer is instructed to execute a set of repetitive tasks quickly and efficiently compared to the processing power of a human" (Lee et al., 2011, p. 33)
Decomposition	"breaking problems down into smaller parts that may be more easily solved" (Barr & Stephenson, 2011, p. 52)
Debugging	"find your own mistakes and fix them" (Hsu et al., 2018, p. 299)
Generalization	"move from specific to broader applicability" (Selby & Woollard, 2013, p. 4)

2.2 Computational Thinking in Higher Education

Literature reviews are available about CT in higher education, in general, and about CT assessment, in particular.

Jong and Jeuring (2020) reviewed 49 articles that examine various interventions aimed at developing CT skills in students. The authors identified the types of interventions, their effectiveness, and how they are evaluated. They found that CT is often taught through programming assignments and that the interventions are evaluated in diverse ways, making comparisons of the findings challenging. Additionally, they noted that interventions are rarely adapted to students' actual proficiency levels. The authors suggest using standardized instruments for evaluating effectiveness and better aligning interventions with students' proficiency levels.

Lyon and Magana (2020) reviewed 13 studies. They analyzed the pedagogical designs and research methods used. The review highlights the need for more research on how students get involved in CT processes or how teaching can improve CT in undergraduate students. Many of the reviewed studies report the positive effects of interventions used to foster CT. However, these studies regularly rely on self-reported metrics. The authors stress the importance of robust definitions and sound measurement instruments to evaluate CT interventions in higher education. Their work emphasizes the growing interest in CT research in higher education.

Lu et al. (2022) reviewed 33 studies about CT assessment in higher education. Most studies targeted undergraduate students, particularly those majoring in computer science. Other groups include in-service teachers and students from various STEM fields, indicating a strong emphasis on technical disciplines. They categorized CT assessment types, including block-based assessments, knowledge/skill tests, self-assessments, text-based programming projects, and academic achievements in computer science courses. Interviews and observations were also utilized to assess CT. The review highlighted challenges in distinguishing between computational thinking skills and computer science skills, with some studies confounding the two. The review identified a need for CT assessments in higher education with an emphasis on research that develops validated assessment instruments.

Zhang et al. (2024) performed a meta-review comprising 11 literature reviews about CT assessment. The major finding of this study is that there is an increase in the number of studies about CT assessment in higher education. A plethora of constructs is available to be included in CT assessments. However, no systematic investigation of the tools or instruments exists regarding the nature of the constructs being assessed. Although there is an (increasing) interest in CT assessment in higher education, issues remain about the methodological rigor and systematic evaluation of interventions.

2.3 Computational Thinking Tests

Shute et al. (2017) point to the Computational Thinking test (CTt) as an internationally established standardized CT assessment instrument. Román-González et al. (2017, 2018) have developed and validated a CTt for secondary students. The authors rely on the CT framework of Brennan and Resnick (2012). It is consistent with the core CT facets (see Table 1). *Abstraction* is covered as visual code blocks representing the problems, including conditionals and variables. *Algorithmic thinking* is necessary because all tasks require sequencing steps to come to a solution. *Automation* is captured by means of loops. *Decomposition* manifests itself in the use of functions to split up the problems into more manageable elements. *Debugging* is necessary because students are required to identify mistakes in provided sequences of code blocks. *Generalization*, however, is not directly addressed by the CTt. For assembling visual code blocks, the authors utilize the *code.org* platform (https://code.org/), which is similar to Scratch (Hsu et al., 2018, p. 302). The test comprises 28 selected response items, can be taken online, and takes about 45 minutes. No programming experience is necessary, which makes this CTt a very flexible instrument. The authors validated their CTt using a sample of 1,251 Spanish 5^{th} to 10^{th} grade students and classical test theory. The reliability of the test is sufficiently high (Cronbach's alpha = .79). Evidence for the suitability of pre-post designs is available (Zhao & Shute, 2019).

In addition to the large amount of time required for the CTt, its main drawback may be that the items are too easy for higher education students. Guggemos et al. (2019) developed a version of the CTt, comprising 26 items, geared towards upper-secondary level students. Five easy items of the initial CTt version were replaced by five more difficult ones. The authors validated the test using 202 students from upper secondary education. They demonstrated Rasch scalability of their CTt version (including unidimensionality) and reported a WLE-reliability of 0.81. Moreover, a proficiency level model was presented, with which the difficulty of items can be predicted. This facilitates the development of new items with a desired difficulty based on validated difficulty drivers (e.g., conditionals and functions). Evidence for the suitability of the test

for longitudinal studies is available (Guggemos, 2021). Although this test yields compelling psychometric characteristics, several items may be too easy for higher education students. This is problematic because for a short CT test, a good fit between proficiency and item difficulty is important (Bühner, 2011).

Lafuente Martínez et al. (2022) developed and validated an Algorithmic Thinking Test for Adults (ATTA) following internationally accepted standards (AERA et al., 2014). The ATTA covers core CT facets (see Table 1): Algorithmic thinking, decomposition, abstraction, pattern recognition, and debugging. It is suitable for assessing people with or without content knowledge about computer science and coding skills in post-secondary education settings. The required test time is about 70 minutes. Cronbach's alpha equals .84. The authors also performed confirmatory factor analyses and concluded the unidimensionality of the ATTA. A drawback of the ATTA may be that it is overall too difficult for undergraduate students in fields outside of STEM. Moreover, the required test time of about 70 minutes for the full version may be a drawback in tertiary education settings.

In summary, combining items from the CTt version for upper-secondary level students (Guggemos et al., 2023) with items from the ATTA may be a good approach for compiling a short test for undergraduate students. The combination may be possible because both tests cover implicitly (CTt) and explicitly (ATTA) the core CT facets (see Table 1). Selecting items from both tests might yield a set of items with a good fit between item difficulty and student proficiency.

3. METHOD

3.1 Compiling the Test

Both the CTt and ATTA have been shown to be unidimensional. Hence, on the content level, all items from either the CTt or ATTA would be suitable for a short CT test. However, the item that elicits the most information about a student's proficiency has an expected probability of a correct answer of 50% (Bühner, 2011). In the Wright map, this is represented by the same Logit value for a person as well as an item. The authors screened the CTt and ATTA items and their reported difficulty (Guggemos et al., 2023; Lafuente Martínez et al., 2022). In combination with data on drivers for item difficulty in CT (Guggemos et al., 2023), a good ex-ante approximation of item difficulty may be possible. Since student CT proficiency substantially varies, a sufficiently broad spectrum of items is necessary. In Rasch modeling, Logit values usually range between ± 2 Logit (Boone, 2016). Using a step of 0.25 Logit, in an optimal case, 17 items would cover the entire proficiency spectrum (± 2 Logit). Overall, we selected 18 items based on predicted item difficulty and predicted student CT proficiency.

3.2 Sample and Data Collection

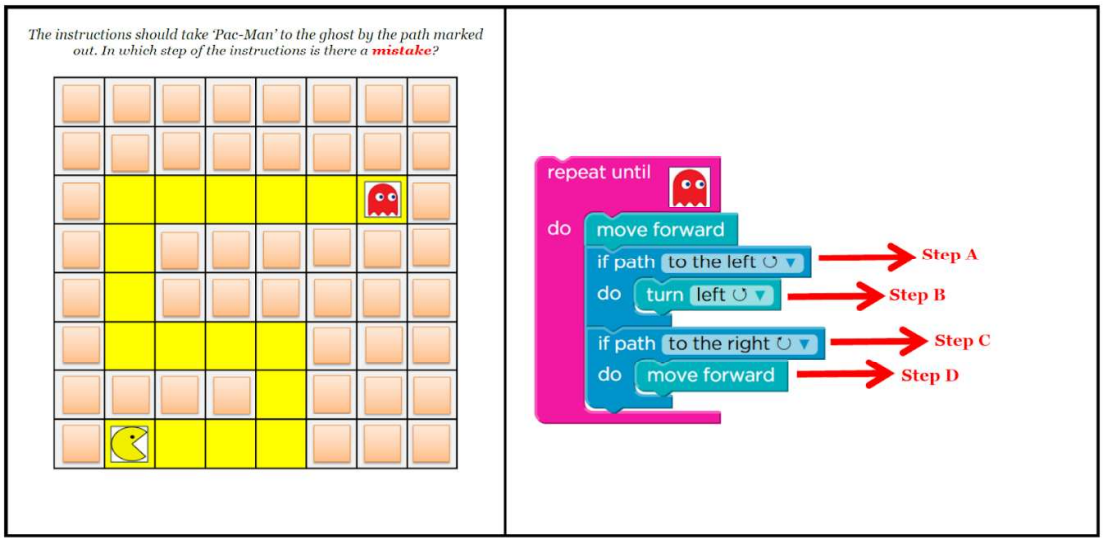
Overall, 290 second-year students from the course "Fundamentals and Methods of Computer Science for Business Studies" at the University of St.Gallen act as a sample. The sample size is sufficient for Rasch modeling (Şahin & Anıl, 2017). The course aims to provide students with foundational knowledge and practical skills in computer science, covering programming, databases, web applications, networking, data science, and machine learning. It follows a structured setup that includes graded weekly quizzes, assignments, project work, and close tutorial support.

Data collection was performed at the beginning of the semester. Prior to performing the test, we collected data on the characteristics of the students. On average, the students were 23.34 years old (SD = 2.22, min = 20, max = 32), and 40% identified themselves as female. We also asked the participants to self-assess their skills in various areas on a ten-point scale: computer skills (M = 4.76, SD = 2.10), programming skills (M = 2.36, 1.86), mathematical skills (6.00, SD = 1.84), English skills (7.65, SD = 1.51), and native language skills (M = 9.11, SD 1.10). During the course's first lecture, the students were asked to perform the test. The lecturer supervised the students and ensured an adequate test environment. The test time was 20 minutes.

To collect the data, we used our own developed multiple-choice assessment and feedback application (LOOM), which over 2000 students have used over the past five years (Rietsche et al., 2018; Ritz et al., 2023). We selected this app because it provides an integrated end-to-end process with a high-quality user

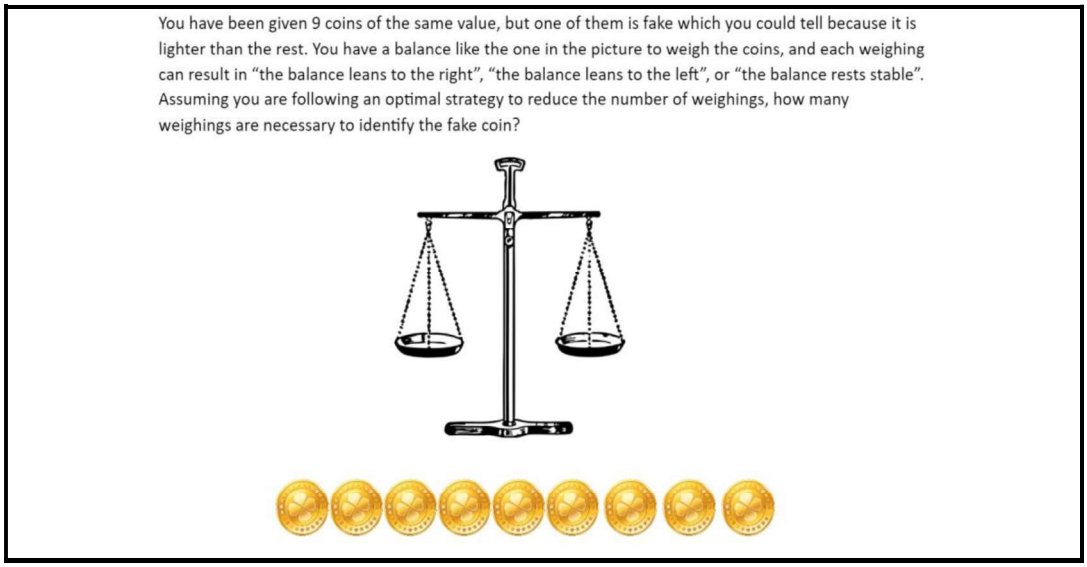
experience for various test items.

We used two testlets. Students were randomly assigned to the testlets (n_1 = 149, n_2 = 141). Each group performed seven unique items, and four anchor items that were identical in both groups, resulting in 11 items per student and, overall, 18 items for both groups (7 + 7 + 4). The items were presented to the students in random order. The anchor items permit locating all the students and items on one logit scale (Boone, 2016). All items are in English and are available here: https://tinyurl.com/yhf6jca8. Figure 1 shows the easiest item (item 14) and Figure 2, the most difficult one (item 18).



Correct answer: Step D.

Figure 1. Item 14 from the test (taken from the CTt)



Correct answer: 2 weighings.

Figure 2. Item 18 from the test (taken from the ATTA)

3.3 Psychometric Test Validation

Due to the favorable characteristic of specific objectivity, we aim at Rasch modeling. This aligns with Zhang et al. (2024), who call for more rigor in research about CT assessment. The main advantage of Rasch modeling is that students (proficiency) and items (difficulty) can be located on a common (Logit) scale that allows for a criterion-referenced test interpretation (Hartig & Frey, 2013). For assessing Rasch scalability and the quality of our test, we rely on the recommendations of Boone (2006). We use a Wright map to evaluate the fit between item difficulty and student CT proficiency. Based on differential item functioning (DIF) analyses, we identify items that may discriminate specific sub-groups (AERA et al., 2014). We use gender (male/female) and the median of students' self-reported skills, e.g., mathematical and programming skills, as split criteria. Based on the median of each self-assessed skill, we form two groups: above-median students and the remainder. We regard DIF less than .43 Logit as negligible, between .44 and .64 as moderate, and greater than .65 as large (Penfield & Algina, 2006).

To check for item homogeneity (unidimensionality), we conduct confirmatory factor analysis ('lavaan' 0.6-18 package in R, Rosseel, 2012) with CT as a single factor. We use a robust maximum likelihood estimator. This, compared to a weighted least squares estimator, allows us to consider missing data by design. The following values serve as cut-off values (van de Schoot et al., 2012): acceptable fit: CFI and TLI >0.90, RMSEA <0.08, SRMR <0.10; good fit: CFI and TLI >0.95, RMSEA <0.05, SRMR <0.06. Since we combined two instruments, we specifically check if a two-dimensional model with the CTt and ATTA as factors fits the data better than a unidimensional model. To compare the competing models, we use the Haughton Bayesian information criterion (HBIC) and the SPBIC variant that is based on a scaled unit information prior and hence more general than the BIC (Lin et al., 2017).

After having checked Rasch scalability, we examine if the items meet the cut-off values applied in the PISA studies (OECD, 2017, pp. 131–134) using the R package 'TAM 4.2-21' (Robitzsch et al., 2024): The deviance from the item discrimination implied by the Rasch model is evaluated utilizing weighted mean square error (wMNSQ = Infit). It should lie between 0.8 and 1.2. The point-biserial correlation should be above 0.30. The percentage of correct answers should fall between 20% and 90%. Not more than 10% of missing data should be present.

Finally, we provide EAP/PV- and WLE-reliability as a measure for overall test reliability from the item response theory context. To ensure comparability with available test instruments, we also report Cronbach's alpha and McDonald omega total from classical test theory; these values should be greater than .70 (Sarstedt et al., 2023).

4. RESULTS

4.1 Psychometric Validity

Table 2 summarizes the characteristics of the test items. The standard deviation of the students' CT proficiency equals 1.60 Logit (mean standardized to zero), with a minimum of -3.61 and a maximum of 3.97 Logit. The mean item difficulty is 0.32 Logit (SD = 1.09). This, and the Wright map (see Figure 3), point to a good fit of item difficulty and CT proficiency. Concerning DIF effects, only negligible DIF appears for gender. However, there are, in parts, large DIF effects in terms of students' self-assessed skills.

The assumption of item homogeneity (unidimensionality) of the test is justified: SB–$\chi^2(144) = 48.67$ ($p = .291$), CFI = 0.984, TLI = 0.980, RMSEA = 0.026, SRMR = 0.049. The non-significant χ^2-test and the fit values point to a good fit. This might also indicate local stochastic independence of the items. A two-dimensional model with the CTt and ATTA as factors fits worse than the one-dimensional model (when considering model complexity): SPBIC = 2,487 and HBIC = 2,417 vs. SPBIC = 3,693 and HBIC = 3,579. Moreover, the latent correlation between the ATTA and CTt equals .941, i.e., these factors can hardly be empirically separated.

Concerning the cut-off values from the PISA studies, in general, all items show good values (see Table 2). The wMNSQ lies between 0.85 and 1.20. The point-biseral correlations are higher than 0.30, except for item 18 (see Figure 2), which is by far the most difficult item. The percentage of correct answers for the

items lies between 0.759 and 0.215, again except for item 18 (0.121). Moreover, item 7 is also slightly too difficult (0.195). Every student fully processed the items, i.e., the proportion of missing values is smaller than 10%. EAP/PV-reliability equals 0.77, and WLE-reliability 0.69. For the first testlet, Cronbach's alpha equals .77 and McDonald's omega total .80. For the second testlet, the values are .75 and .78, respectively. Hence, a short test with 11 items may yield a sufficient reliability for research purposes (>.70).

Table 2. Item difficulty, fit, and DIF-effects of the test (testlet design with two testlets with $n_1 = 141$ and $n_2 = 149$)

						DIF effect					
Item	θ	s.e. θ	wMNSQ	Pt.bis.	P+	Female	Comp.	Prog.	Native	English	Math
1	1.105	0.200	1.002	0.485	0.309	-0.304	0.045	**-1.042**	0.162	-0.275	0.019
2	0.064	0.190	0.916	0.616	0.497	-0.035	-0.028	0.300	-0.355	-0.094	0.093
3	-0.412	0.194	1.016	0.574	0.584	-0.110	-0.376	-0.206	0.247	*-0.596*	**0.919**
4	0.607	0.192	1.116	0.449	0.396	0.366	0.298	-0.421	-0.273	-0.172	-0.058
5*	-0.866	0.145	0.946	0.616	0.662	-0.251	0.128	**0.697**	-0.098	0.010	0.225
6*	0.273	0.136	0.978	0.560	0.459	0.264	0.431	0.300	-0.217	*-0.584*	-0.316
7	1.871	0.228	1.201	0.273	0.195	0.197	0.383	-0.306	-0.343	-0.107	-0.252
8	1.720	0.221	1.016	0.437	0.215	-0.047	0.038	**-0.629**	0.348	*-0.689*	*-0.457*
9	0.461	0.191	1.093	0.484	0.423	-0.432	-1.517	*-0.485*	-0.196	-0.268	**0.677**
10*	-1.290	0.154	0.845	0.684	0.728	0.084	-0.340	0.117	0.160	0.160	-0.341
11*	-0.231	0.138	1.006	0.565	0.552	-0.089	**0.850**	0.234	0.318	-0.316	0.211
12	-0.275	0.198	1.049	0.553	0.560	0.205	-0.385	0.174	*0.521*	**-0.715**	0.275
13	-0.555	0.202	0.862	0.662	0.610	0.224	*0.529*	0.428	*-0.557*	**-0.889**	-0.408
14	-1.516	0.230	0.845	0.666	0.759	-0.020	-0.141	**-0.896**	-0.216	**0.668**	*-0.490*
15	0.733	0.200	1.053	0.504	0.376	0.308	0.427	*0.518*	**0.986**	0.119	0.168
16	0.498	0.197	1.099	0.482	0.418	0.300	-0.166	-0.147	*0.480*	0.005	**-1.020**
17	0.895	0.202	1.052	0.474	0.348	-0.266	-0.250	0.165	**-1.364**	**-0.857**	*-0.530*
18	2.583	0.281	1.157	0.253	0.121	-0.395	0.075	**1.199**	0.395	**4.599**	**1.287**
Min	-1.516	0.136	0.845	0.253	0.121	-0.432	-1.517	-1.042	-1.364	-0.889	-1.020
Max	2.583	0.281	1.201	0.684	0.759	0.366	0.850	1.199	0.986	4.599	1.287

Note. Items in bold are from the ATTA (7-11, 16-18) and all other items are from the CTt. Items marked with an asterisk are anchoring items. DIF effects in italic are moderate, DIF effects in bold are large.
θ = difficulty, s.e. = standard error, wMNSQ = weighted mean square error, Pt.bis. = point biserial correlation, P+ = correct responses.

5. DISCUSSION

This research addresses a critical gap in CT research within higher education, particularly for non-STEM students, by developing an efficient assessment environment. By doing so, this study contributes to the better assessment and tracking of factors that are relevant in enhancing student employability and participation in an increasingly digital society. To this end, we used items from the CTt that is aimed at (upper) secondary level students, as well as the ATTA that addresses adults.

The students in the sample show a very broad proficiency spectrum ranging from -3.61 to 3.97 Logit. Usually, Logit values range between ± 2 Logit (Boone, 2016). This shows the challenge to develop a test with items of suitable difficulty for all students. To illustrate the range: An item that students with medium proficiency (Logit = 0) solve with an expected probability of 50% is solved by the top performers with an expected probability of 98%. The worst-performing students solve this item with an expected probability of 3%. As performance cannot be manipulated by the test takers towards a higher test score (providing a valid

test procedure), the top-performing students may be indeed highly proficient in CT. The finding of a small proportion of students highly proficient in CT is consistent with the ICILS among secondary level students (Fraillon et al., 2019). The low-performing students may not necessarily be poor in CT. Since we tested the students in a low-stakes environment, a lack of test motivation may also explain their poor results (Simzar et al., 2015).

As the Wright map indicates, our set of 18 items meets the students' proficiency spectrum reasonably well. Except for the very high and very low performing students (>|2.0| Logit) there is, in general, an item included that could be solved with an expected probability of 50%. Hence, we were successful in developing an efficient test. Further items could be developed with an expected difficulty greater than 3.5 Logit to precisely measure CT among students with very high CT proficiency. To this end, further items from the ATTA could be used. More importantly, items may be necessary to precisely measure the CT of students with low CT proficiency (<-2.0 Logit). To this end, further items from the CTt can be used. The Wright map can also be utilized to set proficiency standards (Guggemos et al., 2023). If the standard, for instance, is 2.0 Logit, it may not be necessary to measure CT above this level with high precision; rather, in this case, it may be sufficient to ascertain that students are above this level.

Since there is not much room for improvement in terms of person-item fit, adaptive testing may be a viable option to achieve a desired precision with less test time (Chang, 2015). In this case, the item set is not tailored to a specific group but to the individual student. Our research may be a first step towards adaptive testing of CT as the test is, in general, Rasch scalable, which implies specific objectivity. However, considerably more items would be required for adaptive tests than is currently the case.

Logit	Student	Items
4.0		
	XXX	
3.5		
3.0		
2.5	XXXXXXXXX	**18**
2.0		
	XXXXXXXX	**7**
1.5	XXXXXXXXXX	**8**
	XXXXXXXXXXXXXXXXXXX	
1.0	XXXXXXXXXXXXXXXXXXX	1
	XXXXXXXXXXXXXXXXXXX	17
0.5	XXXXXXXXXXXXXXXXXXX	4 15
	XXXXXXXXX	6 **9 16**
0.0	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	2
	XXXXXXXXXXXXXXX	**11**
-0.5	XXXXXXXXXXXXXXX	3 12
	XXXXXXXXXXX	13
-1.0	XXXXXXXXX	5
	XXXXXXXXX	
-1.5	XXXXXXX	**10**
	XXXX	14
-2.0		
	XXXXX	
-2.5	XXXXX	
-3.0		
-3.5	XXXXXXXXXXXXXXXX	
	XXXXXXXXXXXX	
-4.0		

Figure 3. Wright map. Items in bold are from the ATTA; all other items are from the CTt

Concerning the test evaluation, the remarkable DIF effects have to be mentioned. In particular, the DIF effect of item 18 for students with above-median self-reported English skills is noteworthy (DIF = 4.6 Logit). Item 18 is the most difficult one (2.6 Logit). No student with above-median self-reported English skills solved this item correctly, which explains the high DIF effect. The high difficulty of item 18 can also explain the DIF effects regarding self-reported programming and mathematical skills. Further, there is only negligible DIF in terms of gender. This is important because gender differences play an essential role in CT research (Torres-Torres et al., 2024).

Concerning the cut-off values applied in the PISA studies, the items are (almost) within range, except for item 18. The reason for this may be the high item difficulty in combination with the low item discrimination. High item difficulty is, in general, negatively associated with item discrimination (Bühner, 2011). Item 18 could have been removed. However, against the backdrop of the broad proficiency spectrum in our sample, we opted against removing this item.

As demonstrated, the CTt and ATTA are compatible. A unidimensional model yielded the best fit. Both tests are freely available (for research purposes), which allows researchers to cover a very broad CT proficiency spectrum by combining these two tests. This may be important for studies of CT development across the lifespan. Our research therefore opens up avenues for investigating the long-term impact of CT interventions and educational strategies, contributing to evidence-based practices in education. Our test may be particularly suitable for this endeavor as it does not require specific prior knowledge, such as a programming language.

Having demonstrated Rasch-scalability of the combined CTt and ATTA test, the proficiency level model of Guggemos et al. (2023) could be extended to higher education. Such a proficiency level model could help curriculum development and to communicate test results in higher education. Concerning curriculum development, it could be specified what kind of tasks with what cognitive operations students should be systematically able to master after an intervention. When communicating the test results, it could be stated the kind of cognitive operations the student is not able to master but which are required by the curriculum.

Regarding the CTt, Román-González et al. (2018) demonstrated the predictive validity for academic performance and instructional sensitivity. For the ATTA, such evidence is not available. Future research should investigate the predictive validity and instructional sensitivity of the ATTA or those of our test.

Although our test may be suitable for many purposes, we agree with Román-González et al. (2019) that only a combination of instruments may yield a comprehensive picture of CT. Our test may be complemented with self-assessment instruments that capture self-efficacy and can capture CT on a broader scale. Such an instrument may be the Computational Thinking Scales (see Guggemos et al., 2023).

6. CONCLUSION

Valid and economic test instruments are important for formative and summative assessment. We developed a short test for undergraduate students' computational thinking (CT). To this end, we combined the Computational Thinking test and the Algorithmic Thinking Test for Adults. Both performance tests cover generally accepted core facets of CT. Our test shows, in general, good psychometric properties, both in terms of classical test theory and item response theory. The fit between item difficulty and students' CT proficiency is good. The only weakness is some differential item functioning (DIF) effects concerning students' self-assessed skills, e.g., in mathematics. However, there is no substantial DIF concerning gender. All items are selected response and can, therefore, be coded in an economic and objective way. Overall, it may be a suitable instrument to test undergraduate students' CT skills with a reasonable amount of test time (about 20 minutes). Internal consistency reliability (Cronbach's alpha and McDonald's omega total) is greater than .75. In sum, our test may be valuable for research purposes as CT is an important 21st-century skill and may be considered in a variety of research designs. Moreover, our test can be used to ascertain the level of students' CT at the beginning of a course. The obtained information can help lecturers to tailor the course content and offer support to specific students.

REFERENCES

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://eric.ed.gov/?id=ED565876

Barr, V., & Stephenson, C. (2011). Bringing computational thinking to K-12. *ACM Inroads*, *2*(1), 48–54.

Boone, W. J. (2016). Rasch Analysis for Instrument Development: Why, When, and How? *CBE Life Sciences Education*, *15*(4). https://doi.org/10.1187/cbe.16-04-0148

Brennan, K., & Resnick, M. (2012). New frameworks for studying and assessing the development of computational thinking. *American Educational Research Association Meeting, Vancouver, BC, Canada*, 1–25.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion [Introduction to test and questionnaire construction]* (3rd ed.). Pearson Studium.

Chang, H.-H. (2015). Psychometrics behind Computerized Adaptive Testing. *Psychometrika*, *80*(1), 1–20. https://doi.org/10.1007/s11336-014-9401-5

Cheng, M., Adekola, O., Albia, J., & Cai, S. (2022). Employability in higher education: a review of key stakeholders' perspectives. *Higher Education Evaluation and Development*, *16*(1), 16–31. https://doi.org/10.1108/HEED-03-2021-0025

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2019). *Preparing for life in a digital world: IEA International Computer and Information Literacy Study 2018 international report*. International Association for the Evaluation of Educational Achievement (IEA). https://www.iea.nl/publications/study-reports/preparing-life-digital-world

Guggemos, J. (2021). On the predictors of computational thinking and its growth at the high-school level. *Computers & Education*, *161*, 104060. https://doi.org/10.1016/j.compedu.2020.104060

Guggemos, J., Seufert, S., & Román-González, M. (2019). Measuring computational thinking - adapting a performance test and a self-assessment instrument for German-speaking countries. *Proceedings of the 16th International Conference Cognition and Exploratory Learning in the Digital Age (CELDA)*, 183–191. https://doi.org/10.33965/celda2019_201911L023

Guggemos, J., Seufert, S., & Román-González, M. (2023). Computational Thinking Assessment – Towards More Vivid Interpretations. *Technology, Knowledge and Learning*, *28*, 539–568. https://doi.org/10.1007/s10758-021-09587-2

Hartig, J., & Frey, A. (2013). Benefits and limitations of modeling competencies by means of Item Response Theory (IRT). *Zeitschrift Für Erziehungswissenschaft*, *16*(S1), 47–51. https://doi.org/10.1007/s11618-013-0386-0

Hsu, T.-C., Chang, S.-C., & Hung, Y.-T. (2018). How to learn and how to teach computational thinking: Suggestions based on a review of the literature. *Computers & Education*, *126*, 296–310. https://doi.org/10.1016/j.compedu.2018.07.004

Ifenthaler, D., & Yau, J. Y.-K. (2020). Utilising learning analytics to support study success in higher education: a systematic review. *Educational Technology Research and Development*, *68*(4), 1961–1990. https://doi.org/10.1007/s11423-020-09788-z

ISTE. (2015). *Computational thinking: leadership toolkit*. https://www.iste.org/computational-thinking

Jong, I. de, & Jeuring, J. (2020). Computational Thinking Interventions in Higher Education. In N. Falkner & O. Seppala (Eds.), *Koli Calling '20: Proceedings of the 20th Koli Calling International Conference on Computing Education Research* (pp. 1–10). ACM. https://doi.org/10.1145/3428029.3428055

Lafuente Martínez, M., Lévêque, O., Benítez, I., Hardebolle, C., & Zufferey, J. D. (2022). Assessing Computational Thinking: Development and Validation of the Algorithmic Thinking Test for Adults. *Journal of Educational Computing Research*, *60*(6), 1436–1463. https://doi.org/10.1177/07356331211057819

Lee, I., Martin, F., Denner, J., Coulter, B., Allan, W., Erickson, J., Malyn-Smith, J., & Werner, L. (2011). Computational thinking for youth in practice. *ACM Inroads*, *2*(1), 32. https://doi.org/10.1145/1929887.1929902

Lin, L.-C., Huang, P.-H., & Weng, L.-J. (2017). Selecting Path Models in SEM: A Comparison of Model Selection Criteria. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(6), 855–869. https://doi.org/10.1080/10705511.2017.1363652

Lu, C., Macdonald, R., Odell, B., Kokhan, V., Demmans Epp, C., & Cutumisu, M. (2022). A scoping review of computational thinking assessments in higher education. *Journal of Computing in Higher Education*, *34*(2), 416–461. https://doi.org/10.1007/s12528-021-09305-y

Lye, S. Y., & Koh, J. H. L. (2014). Review on teaching and learning of computational thinking through programming: What is next for K-12? *Computers in Human Behavior*, *41*, 51–61. https://doi.org/10.1016/j.chb.2014.09.012

Lyon, J. A., & Magana, J. A. (2020). Computational thinking in higher education: A review of the literature. *Computer Applications in Engineering Education*, *28*(5), 1174–1189. https://doi.org/10.1002/cae.22295

OECD. (2017). *PISA 2015 Technical Report*. OECD Publishing.

Penfield, R. D., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement*, *43*(4), 295–312. https://doi.org/10.1111/j.1745-3984.2006.00018.x

Rietsche, R., Duss, K., Persch, J. M., & Söllner, M. (2018). Design and Evaluation of an IT-Based Formative Feedback Tool to Foster Student Performance. In *Thirty Ninth International Conference on Information Systems (ICIS),* San Francisco, CA, USA.

Ritz, E., Rietsche, R., & Leimeister, J. M. (2023). How to Support Students' Self-Regulated Learning in Times of Crisis: An Embedded Technology-Based Intervention in Blended Learning Pedagogies. *Academy of Management Learning & Education*, *22*(3), 357–382. https://doi.org/10.5465/amle.2022.0188

Robitzsch, A., Kiefer, T., & Wu, M. (2024). *Package 'TAM'*. https://cran.r-project.org/web/packages/TAM/TAM.pdf

Román-González, M., Moreno-León, J., & Robles, G. (2019). Combining assessment tools for a comprehensive evaluation of computational thinking interventions. In S. C. Kong & H. Abelson (Eds.), *Computational thinking education* (pp. 79–98). Springer. https://doi.org/10.1007/978-981-13-6528-7_6

Román-González, M., Pérez-González, J.-C., & Jiménez-Fernández, C. (2017). Which cognitive abilities underlie computational thinking? Criterion validity of the Computational Thinking Test. *Computers in Human Behavior*, *72*, 678–691. https://doi.org/10.1016/j.chb.2016.08.047

Román-González, M., Pérez-González, J.-C., Moreno-León, J., & Robles, G. (2018). Can computational talent be detected? Predictive validity of the Computational Thinking Test. *International Journal of Child-Computer Interaction*, *18*, 47–58. https://doi.org/10.1016/j.ijcci.2018.06.004

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Şahin, A., & Anıl, D. (2017). The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory. *Educational Sciences: Theory & Practice*, *17*, 321–335. https://doi.org/10.12738/estp.2017.1.0270

Sarstedt, M., Hair, J. F., & Ringle, C. M. (2023). "PLS-SEM: indeed a silver bullet" – retrospective observations and recent advances. *Journal of Marketing Theory and Practice*, *31*(3), 261–275. https://doi.org/10.1080/10696679.2022.2056488

Schlax, J., Zlatkin-Troitschanskaia, O., Happ, R., Pant, H. A., Jitomirski, J., Kühling-Thees, C., Förster, M., & Brückner, S. (2020). Validity and fairness of a new entry diagnostics test in higher education economics. *Studies in Educational Evaluation*, *66*, 100900. https://doi.org/10.1016/j.stueduc.2020.100900

Selby, C. C., & Woollard, J. (2013). *Computational thinking: The developing definition*.

Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review*, *22*, 142–158. https://doi.org/10.1016/j.edurev.2017.09.003

Simzar, R. M., Martinez, M., Rutherford, T., Domina, T., & Conley, A. M. (2015). Raising the stakes: How students' motivation for mathematics associates with high- and low-stakes test achievement. *Learning and Individual Differences*, *39*, 49–63. https://doi.org/10.1016/j.lindif.2015.03.002

Tang, X., Yin, Y., Lin, Q., Hadad, R., & Zhai, X. (2020). Assessing computational thinking: A systematic review of empirical studies. *Computers & Education*, *148*, 103798. https://doi.org/10.1016/j.compedu.2019.103798

Torres-Torres, Y.-D., Román-González, M., & Perez-Gonzalez, J.-C. (2024). Didactic strategies for the education of computational thinking from a gender perspective: A systematic review. *European Journal of Education*, *59*(2), Article e12640. https://doi.org/10.1111/ejed.12640

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Voogt, J., Fisser, P., Good, J., Mishra, P., & Yadav, A. (2015). Computational thinking in compulsory education: Towards an agenda for research and practice. *Education and Information Technologies*, *20*(4), 715–728. https://doi.org/10.1007/s10639-015-9412-6

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33–35. https://doi.org/10.1145/1118178.1118215

Wing, J. M. (2008). Computational thinking and thinking about computing. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *366*(1881), 3717–3725. https://doi.org/10.1098/rsta.2008.0118

Zhang, X., Aivaloglou, F., & Specht, M. (2024). A Systematic Umbrella Review on Computational Thinking Assessment in Higher Education. *European Journal of STEM Education*, *9*(1), 1–13. https://doi.org/10.20897/ejsteme/14175

Zhao, W., & Shute, V. J. (2019). Can playing a video game foster computational thinking skills? *Computers & Education*, *141*, 103633. https://doi.org/10.1016/j.compedu.2019.103633