European Journal of Science and Mathematics Education, 2025, **13**(4), 289-303
ISSN: 2301-251X (Online)

OPEN ACCESS

Research Article

Assessing computational thinking skills of science and mathematics upper-secondary school students

Eli Rohaeti [1]
0000-0002-0930-732X

Nur Huda [1*]
0000-0002-0503-8855

[1] Universitas Negeri Yogyakarta, Sleman, Yogyakarta, INDONESIA
[*] Corresponding author: nurhuda.2022@student.uny.ac.id

Citation: Rohaeti, E., & Huda, N. (2025). Assessing computational thinking skills of science and mathematics upper-secondary school students. *European Journal of Science and Mathematics Education, 13*(4), 289-303. https://doi.org/10.30935/scimath/17248

ARTICLE INFO

Received: 7 Jul 2024

Accepted: 21 Jul 2025

ABSTRACT

Computational thinking (CT) is a thinking skill developed and integrated into curricula worldwide in recent years. However, limited assessment is one of the challenges in integrating CT skills into the educational curriculum of developing countries such as Indonesia. This study aimed to develop and validate a CT assessment instrument tailored for upper-secondary school students majoring in science and mathematics in Indonesia. The cross-cultural assessment adaptation method was adopted, comprising six stages: translation, synthesis, back-translation, expert committee review, pretesting, and research audit. Twelve experts were involved in the content validation stage to assess the feasibility of the instrument adapted in Indonesia. The validation process was followed by a pilot test with 501 upper-secondary students majoring in science and mathematics (220 female and 281 male). The data collected were analyzed using the Rasch model measurement. The findings showed that all adapted items met the fit based on the Rasch model measurement, except one spatial question item. The instrument demonstrated high item reliability, although person reliability was relatively low, indicating variation in student responses. The average upper-secondary school students majoring in science have good CT skills. Based on the differential item function value, there are two gender-biased items and four age-biased items. This study hopes to contribute to the literature on CT assessment by providing references and alternative tests for researchers and teachers to use in assessing CT in upper-secondary school students.

Keywords: assessment, computational thinking, Rasch model, science and mathematics student, test adaptation

INTRODUCTION

Since it was promoted by Wing (2006), the term computational thinking (CT) has become an exciting research topic until now. Even if we search on Google Scholar with the keywords “computational thinking research” or “computational thinking education,” there are results of around 1,780,000 to 3,170,000 items. CT not only researched in computer science, programming, and math majors, CT seems to be evolving into a cognitive ability researched in other domains such as science (Hurt et al., 2023; Ogegbo & Ramnarain, 2022; Sari et al., 2025), social sciences (Güven & Gulbahar, 2020), and even integrated in foreign language learning or linguistics (Hsu & Liang, 2021; Rahimi & Sevilla-Pavón, 2025; Rottenhofer et al., 2022). CT is also familiar with being included in STEM education (Cantlon et al., 2024; Revana et al., 2021; Tariq et al., 2025), although various challenges are faced in its implementation (Li et al., 2020). The significant expansion of the work domain of CT skills occurs because CT has relevant cognitive constructs when integrated into other domains outside of computer science or programming. It is not wrong to say that CT is a universal skill everyone needs in the 21st century (Li et al., 2020; Moreno-Leon et al., 2018; Wing, 2006).

The integration of CT into the educational curriculum has also been carried out in various countries and levels of education. For example, early school education curriculum in the United States, parts of Europe, Australia, and Asia In addition, CT has also been integrated into the elementary school curriculum (Waterman et al., 2020; Yang & Lin, 2024) and proven effective in improving the quality of learning (Dagiene & Stupuriene, 2016; Kakavas & Ugolini, 2019). In the secondary school curriculum (K-12), CT has also been implemented in several Asia-Pacific countries (So et al., 2020) and been empirically proven to improve students' problem-solving skills (Chytas et al., 2024; Harangus & Kátai, 2020; Wu et al., 2024).

Computational Thinking Skills in Education

Most CT research is carried out in developed countries; this is closely related to the orientation and condition of the educational curriculum in that country. Developed countries that focus on STEM and the technology industry tend to be more interested in CT because it is proven to increase creativity (Tariq et al., 2025; Xu & Zhang, 2021) and students' problem-solving abilities (Ezeamuzie & Leung, 2022; Voskoglou & Buckley, 2012; Yang et al., 2023). Humans indeed need these two skills in this 21st century. The difference in orientation and conditions of the educational curriculum is what causes not many developing countries to be interested in CT (Chagas & Furtado, 2019), even though many still consider that CT is a thinking skill in computer science only (Li et al., 2020). CT from the beginning was defined as a universal skill that needs to be possessed by everyone, not only computer scientists (Moreno-Leon et al., 2018; Wing, 2006). Other challenges that arise from integrating CT into the educational curriculum are the need for more skilled professionals to implement it and student acceptance of CT, which still needs to be improved (Saidin et al., 2021).

This challenge is also present in the education curriculum of developing countries such as Indonesia. Only a few teachers understand CT and how to integrate it into classroom learning. In addition, the Indonesian national curriculum also does not see CT as a skill students need in the 21st century. So, it is not wrong if CT research is rarely carried out in Indonesia. Even if there is, the research is limited to regional or specific educational institutions. For example, research by Putra et al. (2022) focuses on developing CT tasks in the context of Malay culture, CT ability studies from *Bebras* test results in Sumatra (Zamzami et al., 2020), and adaptation of CT scale for senior high school in Indonesia (Huda & Rohaeti, 2024).

There are many challenges in integrating CT into the education curriculum in Indonesia, including assessment. CT skill assessment tends to be underdeveloped (Angeli & Giannakos, 2020) because assessing CT skills based on their aspects is not easy (Ocampo et al., 2024; Weintrop et al., 2021). Few CT assessment instruments can be applied across countries, cultures, and educational curricula. Tang et al. (2020) have synthesized CT assessments in various studies and obtained 96 instruments with various types of assessments, ranging from traditional tests, portfolios, and surveys to interviews. However, not all assessments are freely accessible to other researchers, so their use is limited. CT assessment that can be used for upper-secondary school students is rarely developed, let alone specifically in the science department. Given that CT cannot be assessed arbitrarily, researchers should not choose the wrong instrument that does not follow the target participants.

Based on an intensive study, researchers found one CT assessment that can be applied to upper-secondary school students in general. The assessment is called *Callysto* computational thinking test (CCTt), developed by Cutumisu et al. (2019a, 2019b) in Canada. Although still rarely used, CCTt has proven valid to be used to measure CT skills in several studies, such as those conducted by Tripon (2022), Jin and Cutumisu (2023), Kamak and Mago (2023), and Cutumisu et al. (2022). In this context, CCTt is an enlightening solution to assessing CT skills in upper-secondary school students majoring in science and mathematics in Indonesia.

Assessment of Computational Thinking Skills

Papert (1980) initiated CT in his book Mindstorms: Children, computers, and powerful ideas. The definition of CT has also evolved to date. Even Voogt et al. (2015) state that there is no definite and final definition of CT. Based on the synthesis carried out by Cansu and Cansu (2019), CT will be close to the process of solving problems and finding solutions. So, it is not strange if CT is considered an essential 21st century skill for everyone (Mohaghegh & McCauley, 2016; Tabesh, 2017). CT dimensions are, in fact, also classified variously by various researchers, but the most popular are decomposition, abstraction, pattern recognition, and algorithms (Cansu & Cansu, 2019). In other studies, different dimensions also emerged, such as automation,

Table 1. Synthesis of CT assessment by school-level

No	Target	Assessment name	Developer
1	Kindergarten and primary school (ages 5–12)	TechCheck-K Beginner's CTT ACES The competent CTT CTT for elementary education	Relkin and Bers (2021) Zapata-Caseres et al. (2020) Paeker et al. (2021) El-Hamamsy et al. (2022) Zapata et al. (2024)
2	Secondary school (ages 10–16)	*Bebras* challenge CTT	Bellettini et al. (2015) Gonzalez (2015)
3	Upper-secondary school (ages 16–18)	CTTt *Bebras* challenge	Cutumisu et al. (2019) Bellettini et al. (2015)

simulation, data collection, data representation, problem decomposition, parallelization, analysis, generalization, evaluation, and debugging (Cansu & Cansu, 2019; ESTE & CSTA, 2011; Tang et al., 2020). All dimensions of CT refer to the human process of solving problems creatively and systematically (Romero et al., 2017).

The significant development of CT to date certainly has problems and challenges. One of the considerable challenges of CT development is the limited number of assessments (Angeli & Giannakos, 2020). Indeed, many CT assessments have developed to date, but not all of them can be used freely. In addition, specialized CT assessments for senior upper-secondary school students are still relatively rare compared to assessments for junior upper-secondary and elementary school students (El-Hamamsy et al., 2025; Tang et al., 2020). **Table 1** shows some CT assessments by school level that can be used today (but some assessments are not freely accessible).

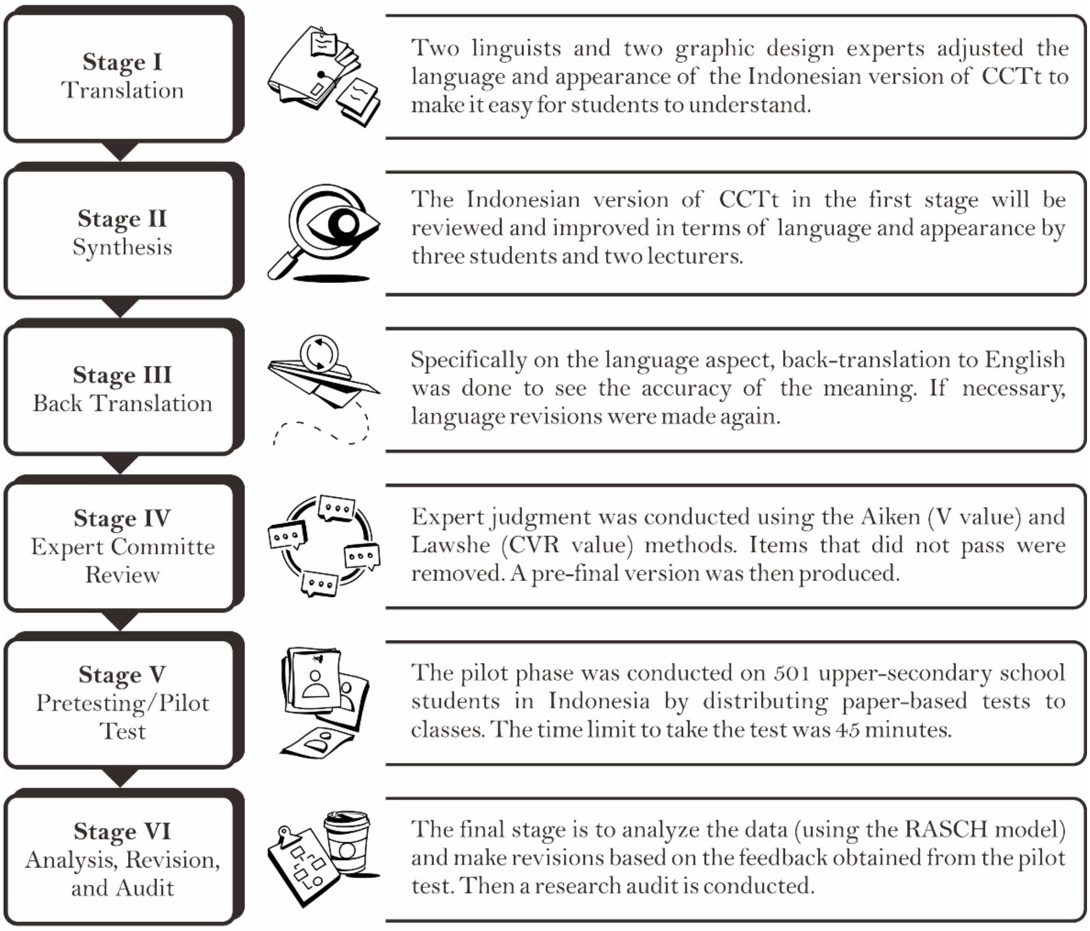
All of the above assessments are organized in a multiple-choice test format. In addition to tests, psychometric assessments in the form of scales or questionnaires have also been developed, such as the computational thinking scales (CTS) (Korkmaz et al., 2017) and the CTS (Ertugrul-Akyol, 2019; Huda & Rohaeti, 2024; X. Li et al., 2024; Tsai et al., 2021). Some even integrate self-efficacy scales with CT (De Jong & Jeuring, 2022). Of the various assessments available, several assessments are developed in the form of tests and questionnaires in one assessment framework, one of which is the CCTt. Cutumisu et al. (2019a, 2019b) and Adam et al. (2019) developed the test in two versions (student version and teacher version). The test was developed in collaboration with 288 students, 150 teachers, and 52 schools in Canada. Implementatively, the CCTt has also been used in several studies, especially to detect the CT ability of pre-service teachers (Jin & Cutumisu, 2023; Tripon, 2022). The validation process of CCTt has been redone more comprehensively with structural equation modelling analysis (Cutumisu et al., 2022). CT research on upper-secondary school students in developing countries such as Indonesia is undoubtedly essential and exciting to do. Unfortunately, assessment limitations hinder the development of CT in Indonesia.

The empirical data will be analyzed using the Rasch model, a measurement model invented by Georg Rasch and applied in education, medicine, and psychology (Andrich, 1988; Snyder & Sheehan, 1992). The Rasch model is mathematically equivalent to the 1PL model in item response theory analysis. Rasch measurement reveals the meaning of students' answers when they take a test or survey questionnaire (Boone, 2016; Yamashita, 2022). The Rasch model's criteria for assessing quality items are based on MNSQ values of 0.6 to 1.4 (Bond & Fox, 2007). Other experts mention that item fit in the measurement model occurs when the MNSQ value is 0.5 to 1.5. Out of this range, the item can be said to be a misfit. Another criterion that can be used to assess item quality is the Z-standard outfit (ZSTD) value in the range of –2.0 to +2.0. However, this ZSTD value tends to be sensitive to large samples (Sumintono & Widhiarso, 2015).

Research Questions

Based on this background, this study aims to adapt and validate the computational thinking test (CTT) for upper-secondary school students majoring in science and mathematics in Indonesia. There are three questions formulated in this study:

1. How is the Indonesian version of the CTT designed?
2. What is the validity and reliability of the CTT based on Rasch model analysis?
3. What is the level of CT skills of science and mathematics upper-secondary school students in Indonesia?



Two linguists and two graphic design experts adjusted the language and appearance of the Indonesian version of CCTt to make it easy for students to understand.

The Indonesian version of CCTt in the first stage will be reviewed and improved in terms of language and appearance by three students and two lecturers.

Specifically on the language aspect, back-translation to English was done to see the accuracy of the meaning. If necessary, language revisions were made again.

Expert judgment was conducted using the Aiken (V value) and Lawshe (CVR value) methods. Items that did not pass were removed. A pre-final version was then produced.

The pilot phase was conducted on 501 upper-secondary school students in Indonesia by distributing paper-based tests to classes. The time limit to take the test was 45 minutes.

The final stage is to analyze the data (using the RASCH model) and make revisions based on the feedback obtained from the pilot test. Then a research audit is conducted.

Figure 1. Cross-cultural adaptation procedure (Source: Authors)

RESEARCH METHOD

Research Design

This study followed the adaptation procedure of the cross-cultural assessment by Beaton et al. (2000). There are six stages, namely:

(1) translation,

(2) synthesis,

(3) back translation,

(4) expert committee review,

(5) pretesting, and

(6) research audit.

With some modifications, the adaptation process will follow the steps shown in **Figure 1**.

Participants

This study involved experts and upper-secondary school students majoring in science and mathematics in Indonesia. At the content validation stage, experts were concerned with reviewing the relevance and feasibility of the test in the context of students in Indonesia. The 12 experts involved in this study included two mathematics education experts (16%), two science and STEM education experts (16%), two educational assessment and evaluation experts (16%), two language experts (16%), two senior math teachers (16%), and two senior science teachers (16%).

Table 2. A table on top

No	Demographic characteristics		Total (n)	Percentage (%)
1	Gender	Male	281	56.1
		Female	220	43.9
2	Age	15	53	10.5
		16	107	21.3
		17	247	49.3
		18	94	18.9
3	Class	X (10)	77	15.3
		XI (11)	88	17.5
		XII (12)	336	67.2
4	Coding experience	Ever	77	15.4
		Never	424	84.6

Which instructions take the "Pacman" (B6) to the "Ghost" (B4) using the path marked out?

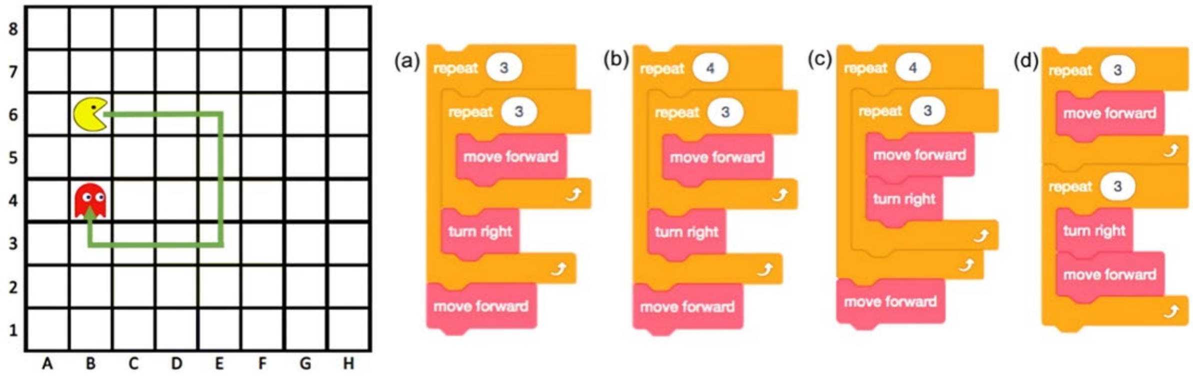


Figure 2. Snapshot of spatial questions from the CTTt (Source: Cutumisu et al., 2019)

In the pilot test phase, 501 upper-secondary school students majoring in science and mathematics in Indonesia were included, consisting of 220 females (44%) and 281 males (56%). **Table 2** shows the demographics of the participants in the pilot test phase. The participants were randomly selected based on classes in schools that permitted the researcher. This study was conducted by submitting an ethical protocol and permission to the original instrument maker.

The target population of this study was students majoring in science and mathematics at public upper-secondary schools in Java Island, Indonesia, with a total of approximately 1 million students. Based on the measurements according to Krejcie and Morgan (1970) and Taherdoost (2017), the number of participants in this study can already represent the population (with notes: variance of the population P = 50%, confidence level = 95%, margin of error = 5%).

Data Collection

The CCTt by Cutumisu et al. (2019a, 2019b) is organized into three constructs that measure students' conceptual understanding, practice, and viewpoints toward CT skills. The three constructs are a polytopic scale questionnaire (n = 19), spatial questions (SQ) in multiple-choice form (n = 11), and problem-solving questions (PSQ) in open-ended format (n = 2). An example of a spatial test in CCTt can be seen in **Figure 2**. This test was developed for students in grades 5–12 and teachers. So, CCTt consists of two versions: the student and teacher versions.

In the last section, there were two PSQs in the form of open-ended questions. Based on the official information, CCTt was prepared without an answer key, so the researcher had to create his answer key to analyze the test results students had done.

Data Analysis

The data in this study are divided into two, namely: content validation data from expert judgment and raw score data from test trials. The analysis calculates Lawshe's CVR index value for questionnaires and Aiken's V for spatial and PSQs in content validation. Lawhse's CVR analysis was conducted with the formula: $CVR =$

$(ne - N/2)/(N/2)$, where ne is the number of experts who answered "essential" and N is the total experts (raters). Aiken's V formula is $V = \sum S/[n(c-1)]$, where $S = r - lo$. c is the highest assessment number, lo is the lowest, and n is the number of raters who assess. The acceptable CVR value based on the number of raters (N = 12) is 0.56 (Lawshe, 1975), while for Aiken's V value with 12 raters, 4 number of rating categories, and 5% significance level is 0.69 (Aiken, 1980). If an item does not meet this value, it will be discarded (not included in the pilot test stage). The content validation in this study also contains more or less qualitative data in the form of input and suggestions from experts on the CCTt adapted into the Indonesian version. The qualitative data will be described and summarized as revisions before the pilot test stage.

Data from the pilot test/pretesting stage were raw scores, then analyzed using the Rasch model measurement. This measurement can overcome the limitations of classical test theory in assessing test quality and explaining the validity of an assessment (Boone et al., 2014). The value calculated in the Rasch model is the probability score (P), which is converted into a logit function (W-score) or logarithmic conversion of the calculation: odd ratio = $P/(1-P)$. The tool used to analyze the scores in the Rasch model in this study is Winsteps version 3.73.

Several critical criteria must be met for an item to be said to fit in the Rasch model, including item fit based on MNSQ and ZSTD values, unidimensional assessment of Rasch PCA, local dependency analysis related to the correlation between items, item polarity based on PTMEA CORR values, and others. The above five criteria are significant limitations that must be analyzed in the Rasch model. Items in an assessment can be said to validly measure what should be measured if they meet the five criteria above. The reliability value in the Rasch model is calculated by the separation reliability value (item or person reliability) with a minimum value of 3.00. Reliability in Rasch can also be determined by the value of item reliability and person reliability with a minimum value of > 0.67 (sufficient category). Internal consistency in Rasch can be calculated based on the value of Cronbach's alpha with a good category if it is more than 0.7 (Nunnally, 1978; Taber, 2018).

RESULTS

Content Validity

Based on the assessment of 12 competent experts, the analysis was carried out using the validity measurement method by Lawshe and Aiken. Lawhse's CVR calculation was carried out to see the content validity of 19 CCTt questionnaire items containing five aspects, namely: digital literacy (DIL), CTT, coding experience (CE), data literacy (DAL), and computational thinking experience (CTE). Based on all aspects assessed, only CE did not meet the valid criteria with an average CVR value = 0.375 below the minimum limit of < 0.56 (Lawshe, 1975). This value can be interpreted as many experts stated that the questionnaire items related to CE were not essential when asked to upper-secondary school students in Indonesia. The content validity index of the CCTt questionnaire was 0.79. The perspectives of educational assessment and assessment experts were the most extreme and significant; even one assessment expert stated that the four questionnaires on CE were not essential to ask, four other experts stated that CE was important but not essential to ask.

Interesting results were also seen in the content validity of multiple-choice question 11 (SQ11), which did not meet the minimum Aiken's V value. SQ11 is a PSQ in CCTt that is closest to programming languages. SQ11 obtained a value of V = 0.58 in the calculation, which did not meet the minimum requirement of > 0.69 (Aiken, 1980). It can be concluded that experts consider problem items and questionnaires very close to coding difficult for upper-secondary school students majoring in science and mathematics in Indonesia because they are not taught in the current curriculum.

Based on the results of the calculation of content validity based on the CVR and V values above, it can be concluded that the CCTt adapted into the Indonesian version was reduced to 15 questionnaires, 10 SQs, and 2 PSQs, as in **Table 3**. The construction was then tested on upper-secondary school students majoring in science and mathematics in Indonesia and analyzed using the Rasch model measurement.

Table 3. Indonesian version of the test construction after content validation by experts

Indicator	Item description
DIL	4 Likert scale items containing statements such as “I find it easy to use technology” and “People ask me for help using their computers”.
CT concept	4 Likert scale items containing statements such as “When I solve a complex problem, I try to break it down into smaller or simpler problems” and “When I solve a complex problem, I think about other problems I have solved before to see if I can solve this problem in the same way”.
DAL	3 Likert scale items with statements such as “I would rather explore the data myself than have someone tell me what it means” and “I get frustrated when trying to understand the data”.
CTE	4 Likert scale items with statements such as “It is important to develop computational thinking” and “I have the skills to teach others about computational thinking”.
SQ	10 multiple choice question items with four options A, B, C, and D
PSQ	2 open-ended questions on “birth of a baby girl” and “finding the lightest coin”

Table 4. Measurement item fit statistics

Item	Measure	Standard error	MNSQ		PT-measure	
			Infit	Outfit	Corr.	Exp.
SQ4	0.56	.11	1.33	1.57	.34	.54
SQ8	0.03	.11	1.12	1.15	.45	.52
SQ7	–0.12	.11	1.11	1.10	.54	.51
SQ5	0.81	.11	0.99	1.07	.55	.55
SQ2	–0.69	.11	0.84	0.97	.55	.47
SQ9	0.81	.11	0.91	0.95	.59	.55
SQ1	–1.31	.12	0.91	0.92	.47	.42
SQ6	0.93	.11	0.91	0.87	.60	.55
SQ10	–0.68	.11	0.90	0.83	.54	.48
SQ3	–0.33	.11	0.87	0.81	.57	.50

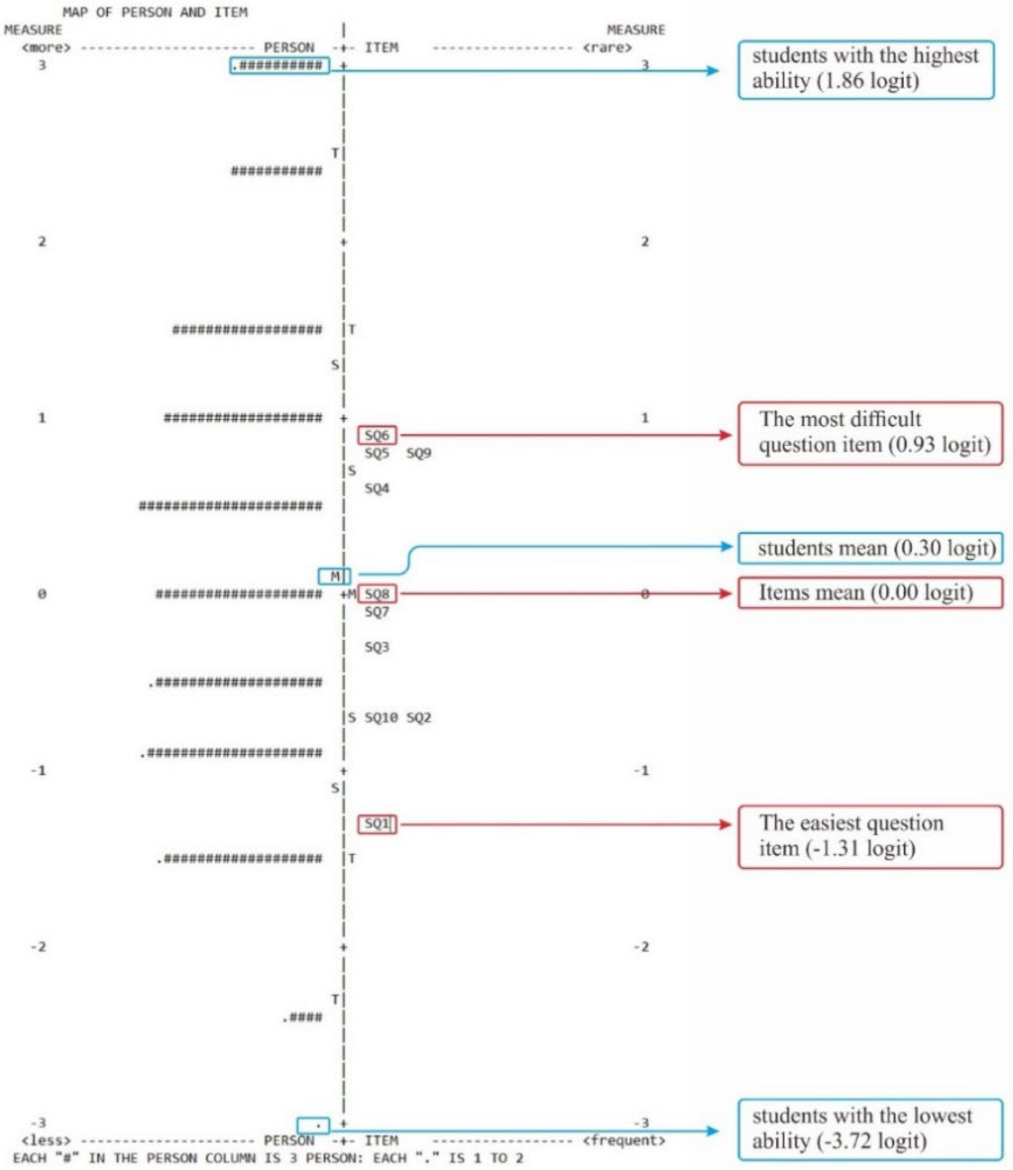
Item Fit Analysis

Item fit in Rasch model measurement tends to be affected by the sample size in a study. However, the general rule of thumb in measuring item fit to the Rasch model refers to the infit (inlier-sensitive or information-weighted fit) and outfit (outlier-sensitive or information-weighted fit) values in MNSQ and ZSTD. The acceptable MNSQ value for model fit is 0.5 ≤ MNSQ ≤ 1.5. **Table 4** shows the results of the calculation of item fit with the Rasch model. It can be seen that all items have MNSQ values in the range of 0.84 < MNSQ Infit < 1.33 and 0.81 < MNSQ Outfit < 1.55. These values indicate that all items in CCTt are valid, except item SQ4, which has an MNSQ outfit value = 1.57. Thus, item SQ4 did not fit in the CCTt pilot in Indonesia. The non-fit of item SQ4 is also confirmed by the outfit ZSTD = 7.0 and Infit ZSTD = 6.5. At the same time, the threshold of acceptable ZSTD value is –2.0 ≤ ZSTD ≤ +2.0 (Bond & Fox, 2007). Therefore, item SQ4 must be revised, retested, or discarded. In this study, SQ4 will be discarded as a misfit item. As for items other than SQ4, all of them are fit based on the Infit and Outfit MNSQ values.

A valid assessment is expected to have item polarity that measures the same construct. Item polarity in Rasch model measurement is analyzed using the point measure correlation (PTMEA CORR) value with a minimum acceptable value of > 0.30 (Wu & Adams, 2007). In other sources, the acceptable PTMEA CORR value is 0.40 < PTMEA CORR < 0.85 (Sumintono & Widhiarso, 2015). **Table 4** shows that all PTMEA CORR values from the CCTt assessment fall into the fit category, except for item SQ4. The PTMEA CORR value of item SQ4 is 0.034. The value is minimal enough to be used as evidence that SQ4 is valid. The PTMEA CORR value of SQ4 also confirms the misfit of the item based on the outfit MNSQ calculation. Based on these values, it can be concluded that the Indonesian version of CCTt has measured the same construct, namely CT skills, except for item SQ4.

Wrigh-Map Analysis

Ideal item map conditions occur when a test item or questionnaire represents each interval in the item map (Sumintono & Widhiarso, 2015). The limits of acceptable difficulty levels are +3.00 to –3.00 (Andrich & Styles, 2004). Based on the CCTt pilot test data, the Rasch model analysis showed that the range of item



```
        MAP OF PERSON AND ITEM
MEASURE                                   |                               MEASURE
  <more> --------------------- PERSON -+- ITEM    ---------------- <rare>
   3                        .##########  +                               3
                                         |
                                         |
                                         |
                                        T|
                            ###########  |
                                         |
                                         |
                                         |
   2                                     +                               2
                                         |
                                         |
                                         |
                                         |
                     ##################  |T
                                         |
                                        S|
                                         |
   1                ###################  +                               1
                                         | SQ6
                                         | SQ5  SQ9
                                         |S
                                         | SQ4
             ######################      |
                                         |
                                         |
                                        M|
   0                 ###################  +M SQ8                          0
                                         | SQ7
                                         |
                                         | SQ3
                                         |
                 .####################   |
                                         |
                                         |S SQ10 SQ2
                                         |
               .#####################    |
  -1                                     +                              -1
                                        S|
                                         |
                                         | SQ1
                                         |
                  .###################   |T
                                         |
                                         |
                                         |
  -2                                     +                              -2
                                         |
                                         |
                                        T|
                                  .####  |
                                         |
                                         |
                                         |
                                         |
  -3                                .    +                              -3
  <less> --------------------- PERSON -+- ITEM    ---------------- <frequent>
EACH "#" IN THE PERSON COLUMN IS 3 PERSON: EACH "." IS 1 TO 2
```

Figure 3. Spatial question Wright map (Source: Authors)

measure values was –1.31 < measure < 0.93 with an average measure value of 0.00 on all items. It can be concluded that the adapted CCTt has a difficulty level equivalent to the ability of upper-secondary school students majoring in science and mathematics in Indonesia.

Figure 3 shows a Wright map that provides a clearer picture of the distribution of student abilities and the difficulty level of the questions in logit units. Based on the person-item map from **Figure 3**, 32 students (6.3%) have the highest ability with a value of 1.86 logits. In addition, 176 students (34.9%) can answer the most challenging question (SQ6). In contrast, two students (0.3%) had the lowest ability, and 74 (14.7%) had the lowest ability and could not answer the most straightforward question. Based on these logit values, more students can do CCTt than those who cannot. So, it is natural that the average logit value of students is 0.30 logit. This value means the students who participated in this study could do the CCTt.

Based on the person-item map above, 32 students (6.3%) have the highest ability with a value of 1.86 logits. In addition, 176 students (34.9%) can answer the most challenging question (SQ6). In contrast, two students (0.3%) had the lowest ability, and 74 students (14.7%) had the low ability and could not answer the most straightforward question. Based on these logit values, more students can do CCTt than those who cannot. So, it is natural that the average logit value of students is 0.30 logit. This value means the students who participated in this study could do the CCTt.

The order of question difficulty from lowest to highest for SQs is SQ1, SQ2, SQ10, SQ3, SQ7, SQ8, SQ4, SQ9, SQ5, and SQ6. SQ1 and SQ2 are the fundamental questions in this instrument, so it is not surprising that the difficulty level is the easiest (–1.31 and –0.69), and 381 students (83%) answered correctly. As SQ6 was the highest difficulty question (0.93 logits), it is not surprising that 310 students (62%) answered incorrectly. However, 191 students (38%) still got it right. As for the PSQs, the measure values were 0.36 logit for PSQ1 and –0.36 logit for PSQ2. The "birth rate of female babies" question was more complicated than the "finding the lightest one coin." The mean logit value for the person on PSQs was –2.65. It can be concluded that students' ability is lower than the difficulty level of the questions. The evidence is that only seven students (1.39%) could answer PSQ1 and PSQ2 correctly. There were 46 students (9.18%) who did not answer PSQ1 and 37 students (7.38%) who did not answer PSQ2. Based on the logit value, PSQs are more challenging than spatial ones.

Differential Item Functioning (DIF)

Biases that need to be detected in adapting CCTt instruments in the Indonesian context are gender bias, age, CE, and test-taking time. A good item should not contain elements of bias and favor one particular group. In other words, items are said to be biased if individuals with the same ability from different groups have different probabilities when answering questions (Kelderman, 1989). The DIF test in Rasch analysis can be seen based on the person DIF plot graph and the probability value of the two-tailed t-test. Items that have a probability value < 0.05 are indicated as biased items and favor one particular group.

Gender bias was detected in items SQ4 and SQ9. Bias caused by participants' age was detected in items SQ1, SQ2, SQ4, and SQ7. Bias caused by processing time was detected in items SQ2, SQ3, and SQ7. In comparison, the bias caused by the CE factor is detected in items SQ2, SQ8, PSQ1, and PSQ2. The item that has the most bias is SQ2. At the same time, the items with no bias are items SQ6 and SQ10. Interestingly, the most challenging item in CCTt is the item with no bias. Participant age was the cause of question bias, with an average p-value of 0.22. Although not below 0.05, this value is lower than other aspects (gender, time, and CE). Based on DIF analysis, age strongly determines students' answers.

Assessment Reliability and Separation Index

Reliability in the Rasch model can be seen based on Cronbach's alpha value, person reliability, and item reliability. An assessment is of high quality and reliable enough to be used when it has an alpha Cronbach reliability value of at least 0.70; this value indicates the internal consistency of an assessment (Nunnally, 1978; Taber, 2018). The value of person and item reliability can be pretty good if it is above 0.67; this value can be interpreted as the consistency of respondents' answers and the consistency of items in an assessment (Sumintono & Widhiarso, 2015). In the CCTt adapted in Indonesian, the Cronbach alpha reliability value obtained is 0.70; this value is in the moderately reliable category. The person reliability value is at a value of 0.63, which is in the weak category. However, the item reliability of the Indonesian version of CCTt is at a value of 0.98, which is in the excellent category. The consistency of the CCTt items in the Indonesian version is high, but the consistency of student answers is relatively low.

DISCUSSION

Content validation has been carried out by calculating Lawshe's CVR and Aiken's V values, resulting in a new CCTt construction consisting of 15 questionnaires, 10 SQs, and 2 PSQs. The new construction slightly differs from the original version of CCTt by Cutumisu et al. (2019a, 2019b) or Adams et al. (2019). An interesting finding from the content validation stage was the low expert ratings of the questionnaire items and questions related to coding. One plausible reason why items close to coding are not essential in the Indonesian version of CCTt is that the average student in Indonesia is not taught coding, as revealed by rater 3:

> The school curriculum in Indonesia does not provide a coding experience in the learning process. Thus, statements in the coding experience aspect will be difficult to answer and are guaranteed to be answered carelessly, carelessly, or guessing, so they should be deleted and not revised.

Table 5. Categorization of student answer types on problem-solving questions

Item	Student answer categories		
	1	2	3
PSQ1	A. Larger cities: Because larger cities have larger areas, the potential birth rate of girls is also greater.	B. Smaller cities: Because smaller cities have small areas, it is easier for small cities to achieve a 60% girl birth rate than large cities.	C. Almost the same: Due to the same probability with different area differences, it is most likely that the birth rate is the same, only 5% difference from each other.
PSQ2	One coin is placed on one side of the scale. Then other coins are tried until you get the lightest coin.	The 10 coins are divided into two parts. 5 coins are placed on one side of the scale, and the other 5 are placed on the other. Of the 5 lighter coins, the lightest coin is found by trial and error.	10 coins are divided into three parts, leaving 1 coin. Then weighed, if the three parts are equal, then the remaining 1 coin is the lightest.

So, it is natural that there is still a lack of research on coding learning or curriculum in Indonesia; in contrast to other countries, countries such as the UK, Australia, Greece, and Estonia have integrated computing skills into their curriculum since 2010 (Popat & Starkey, 2019). Other items unrelated to coding were valid and considered essential and vital to retain, although there were minor revisions. The revisions provided by the experts were related to typography, improvement of working instructions, illustration of pictures, and consistency of terms.

In the redesign process of the Indonesian version of CCTt, the colors of some objects were changed. The goal is to be clear when printed into paper form because the Indonesian version of the CCTt is given in a paper-based test (PBT). Although it is said that PBT and computer-based tests are equivalent in quality (Herrmann-Abell et al., 2018), the researcher considers it more appropriate to administer the Indonesian version of the CCTt with PBT. The aim is to improve the quality of student work, reduce screen distraction, and increase concentration so that students can complete the given questions more quickly (Noyes & Garland, 2008).

Distraction from working on test questions also occurs because students need help understanding instructions. There were unfamiliar terms such as “loop” and “if, then, else” instructions that students did not know. The unfamiliar terms also caused items SQ6 and SQ5 to be the most difficult questions. SQ6 is the type of problem that contains the most if, then, else concepts, so it is complex to solve. The same applies to item SQ5. Besides complex question construction, items SQ6 and SQ5 also have strong distractors. In item SQ6, 257 students (51.2%) answered option C. Compared to option D as the correct answer, the difference between the two options lies only in the number of words “if-then.” Answer D is more complex and contains more “if then” than answer C. Answering D is undoubtedly the most reasonable answer for students who think more deeply. Moreover, in answer C, one loop does not function to move the object. This finding also occurred in item SQ5, where 173 students (34.5%) answered option A incorrectly. The difference between option A and option B as the correct answer is also the same as item SQ6, which is the number of words “if then.” Answer A looks simpler than answer B. The indication is that low-ability students need to think more deeply when answering complex questions.

Another interesting finding is that low-ability students tend to answer the PSQs inadequately or not answer them at all. There were 46 students (9.1%) who did not answer PSQ1, and there were 34 students (6.7%) who did not answer PSQ2. This data proves that PSQ1 tends to be more difficult than PSQ2. When confirmed through the person measure of spatial problem items, it can be concluded that the average student who does not answer the PSQ is a low-ability student. Low-ability students are reluctant to think more deeply about solving problems. Another possibility that PSQs still need to be answered is the factor of insufficient time limit allocation for the test.

The Indonesian version of CCTt is conducted according to the original, 45 minutes [30]. This time is equivalent to one hour of upper-secondary school lessons in Indonesia. More than 45 minutes is needed for students in Indonesia to complete the CCTt. At the last minute, students tend to rush because they still need to complete the PSQs. The author recommends that CCTt be done with a time limit of 60 minutes in future research. So, PSQs can be completed before the time runs out. The types of answers students give on PSQs can be seen in **Table 5**.

Based on the scores of all participants, only 7 students (1.39%) could answer the question correctly (answering as in category 3 answers). In other words, only 7 students had high-level CT skills. The facts from this data also conclude that problem-solving is a type of enrichment problem that is more difficult to solve than multiple-choice spatial problems. The finding has also been confirmed based on the function of the previous measurement information.

Based on the DIF calculation, the Indonesian version of CCTt has a slight gender bias. This finding aligns with previous research, which states that CT ability differs by gender (Niousha et al., 2023; Sovey et al., 2022). Although subtle, this fact must still be revealed. The timing of the test administration also caused bias in some answers. Students who took the CCTt in the morning scored higher than students who took the CCTt during the day. The reason is that during the day, the concentration of students has decreased when working on quite complex problems. So, it is recommended that future research equalize the time of CCTt in the morning. Thus, the results will be more optimal because students still fully concentrate when working. In addition, the age factor also needs to be considered. In some questions, an age bias was detected. One of the reasons is that the scores of 15-year-old students are unnatural or abnormal. There is a slight possibility that students aged 15 tend to guess the answers to complex questions.

Another interesting finding from the CCTt adaptation in Indonesia is answer bias due to the CE factor, especially on PSQs. This finding aligns with research conducted by Zamzami et al. (2020), which states that CE significantly affects students' CT skills. All students with the lowest ability on SQ items and students who did not answer on PSQ items needed CE. Thus, it is recommended that students with no coding knowledge and those with CE be separated when measuring CT skills in Indonesia. After passing a rigorous validation stage with content validation and pilot testing, the Indonesian version of CCTt was reduced to 15 questions, 9 SQs, and 2 PSQs. This construction slightly differs from the English version of CCTt used by other researchers, such as Jin and Cutumisu (2023) and Tripon (2022).

CONCLUSION

This study presents a new CT skills assessment that can be widely used in Indonesia. The Indonesian version of the CTT has been adapted and validated and resulted in an assessment construct consisting of 15 questionnaires, 9 SQs, and 2 PSQs. The Indonesian version of the test items fit the Rasch model measurement and can be used for further research. As a note, it is better to administer the test in the morning within the time limit of 60 minutes so that students' concentration is still optimal for working on the test. Based on the differential item function measurement, some items were gender-biased, CE-biased and age-biased. Therefore, it is recommended in future research to separate the analysis between male and female students, student age, and students who have CE and not when using test as an assessment. However, not all items were detected as biased, so tests can still be used as a comparative assessment of students' CT ability based on various factors or predictors.

The results of this study also show that upper-secondary school students majoring in science and mathematics in Indonesia have good CT thinking skills even though they have never received CT learning. The existence of the Indonesian version of test is expected to make it easier for teachers and researchers to observe and assess students' CT skills in Indonesia more easily and accurately. Thus, efforts to integrate CT in Indonesia are increasingly widespread. This study also hopes to contribute to the literature on CT assessment by providing references and alternative tests for researchers and teachers in assessing CT in upper-secondary school students. Despite the contributions above, this study was limited to upper-secondary school students majoring in science and mathematics. Therefore, future research is recommended to pilot test in other majors besides science, such as mathematics, social studies, or vocational high school majors.

Author contributions: Both authors were involved in concept, design, collection of data, interpretation, writing, and critically revising the article. Both authors approved the final version of the article.

Acknowledgments: The authors would like to thank all the schools and individuals involved for their willingness to participate in this research. This research was supported by the Program of World Class University – Badan Pengelola Perguruan Tinggi Negeri Berbadan Hukum (WCU-BPPTNBH), Universitas Negeri Yogyakarta, 2025.

Ethics declaration: This study was approved by the Ethics Commission of Universitas Negeri Yogyakarta (Approval Code: 00052134041114420230625151; Approval Date: July 21, 2023). All participants provided informed consent prior to participation. Personal data were collected and processed anonymously to ensure confidentiality and privacy in accordance with ethical guidelines.

Declaration of interest: The authors declared no competing interest.

Data availability: Data generated or analyzed during this study are available from the authors on request.

REFERENCES

Adams, C., Cutumisu, M., Yuen, C., Hackman, Lu, C., & Samuel, M. (2019). Callysto computational thinking test (CCTt) teacher version. *Callysto*. https://callysto.ca/computational-thinking-tests/callysto-computational-thinking-test-cctt-teacher-version/Resource_Callysto-CTt_Instrument_Teacher_version-1.pdf

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement, 40*(4), 955–959. https://doi.org/10.1177/001316448004000419

Andrich, D. (1988). *Rasch models for measurement*. SAGE. https://doi.org/10.4135/9781412985598

Andrich, D., & Styles, I. (2004). Final report on the psychometric analysis of the early development instrument (EDI) using the Rasch model. *AEDI*. https://api.semanticscholar.org/CorpusID:141656000

Angeli, C., & Giannakos, M. (2020). Computational thinking education: Issues and challenges. *Computers in Human Behavior, 105*, Article 106185. https://doi.org/10.1016/j.chb.2019.106185

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine, 25*(24), 3186–3191. https://doi.org/10.1097/00007632-200012150-00014

Bellettini, C., Lonati, V., Malchiodi, D., Monga, M., Morpurgo, A., & Torelli, M. (2015). How challenging are Bebras tasks?: An IRT analysis based on the performance of Italian students. In *Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education* (pp. 27–32). ACM. https://doi.org/10.1145/2729094.2742603

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum Associates Publishers.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE—Life Sciences Education, 15*(4), Article rm4. https://doi.org/10.1187/cbe.16-04-0148

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer. https://doi.org/10.1007/978-94-007-6857-4

Cansu, F. K., & Cansu, S. K. (2019). An overview of computational thinking. *International Journal of Computer Science Education in Schools, 3*(1), 17–30. https://doi.org/10.21585/ijcses.v3i1.53

Cantlon, J. F., Becker, K. T., & DeLong, C. M. (2024). Computational thinking during a short, authentic, interdisciplinary STEM experience for elementary students. *Journal for STEM Education Research, 7*, 425–443. https://doi.org/10.1007/s41979-024-00117-0

Chagas, D., & Furtado, E. (2019). Computational thinking in basic education in a developing country perspective. In A. Visvizi, & M. D. Lytras (Eds.), *Research & innovation forum 2019* (pp. 135–150). Springer. https://doi.org/10.1007/978-3-030-30809-4_14

Chytas, C., Van Borkulo, S. P., Drijvers, P., Barendsen, E., & Tolboom, J. L. J. (2024). Computational thinking in secondary mathematics education with GeoGebra: Insights from an intervention in calculus lessons. *Digital Experiences in Mathematics Education, 10*, 228–259. https://doi.org/10.1007/s40751-024-00141-0

Cutumisu, M., Adams, C., Glanfield, F., Yuen, C., & Lu, C. (2022). Using structural equation modeling to examine the relationship between preservice teachers' computational thinking attitudes and skills. *IEEE Transactions on Education, 65*(2), 177–183. https://doi.org/10.1109/TE.2021.3105938

Cutumisu, M., Adams, C., Yuen, C., Hackman, L., Lu, C., & Samuel, M. (2019a). Callysto computational thinking test (CCTt) student version. *Callysto*. https://callysto.ca/computational-thinking-tests/callysto-computational-thinking-test-cctt-student-version/Resource_Callysto-CTt_Instrument_Student_version-1.pdf

Cutumisu, M., Adams, C., Yuen, C., Hackman, M., Lu, C., & Samuel, M. (2019b). Computational thinking test. *Callysto*. https://callysto.ca/computational-thinking-tests/

Dagiene, V., & Stupuriene, G. (2016). Informatics concepts and computational thinking in K-12 education: A Lithuanian perspective. *Journal of Information Processing, 24*(4), 732–739. https://doi.org/10.2197/ipsjjip.24.732

De Jong, I., & Jeuring, J. (2022). Developing a self-efficacy scale for computational thinking (CT-SES). In *Proceedings of the 22nd Koli Calling International Conference on Computing Education Research* (pp. 1–2). https://doi.org/10.1145/3564721.3565954

El-Hamamsy, L., Zapata-Cáceres, M., Martín-Barroso, E., Mondada, F., Zufferey, J. D., Bruno, B., & Román-González, M. (2025). The competent computational thinking test (cCTt): A valid, reliable and gender-fair test for longitudinal CT studies in grades 3–6. *Technology, Knowledge and Learning*. https://doi.org/10.1007/s10758-024-09777-8

Ertugrul-Akyol, B. (2019). Development of computational thinking scale: Validity and reliability study. *International Journal of Educational Methodology, 5*(3), 421–432. https://doi.org/10.12973/ijem.5.3.421

ESTE, & CSTA. (2011). Operational definition of computational thinking for K-12 education. *ISTE*. https://cdn.iste.org/www-root/Computational_Thinking_Operational_Definition_ISTE.pdf

Ezeamuzie, N. O., & Leung, J. S. C. (2022). Computational thinking through an empirical lens: A systematic review of literature. *Journal of Educational Computing Research, 60*(2), 481–511. https://doi.org/10.1177/07356331211033158

Güven, I., & Gulbahar, Y. (2020). Integrating computational thinking into social studies. *The Social Studies, 111*(5), 234–248. https://doi.org/10.1080/00377996.2020.1749017

Harangus, K., & Kátai, Z. (2020). Computational thinking in secondary and higher education. *Procedia Manufacturing, 46*, 615–622. https://doi.org/10.1016/j.promfg.2020.03.088

Herrmann-Abell, C. F., Hardcastle, J., & DeBoer, G. E. (2018). Comparability of computer-based and paper-based science assessments. In *Proceedings of the NARST Annual International Conference*.

Hsu, T.-C., & Liang, Y.-S. (2021). Simultaneously improving computational thinking and foreign language learning: Interdisciplinary media with plugged and unplugged approaches. *Journal of Educational Computing Research, 59*(6), 1184–1207. https://doi.org/10.1177/0735633121992480

Huda, N., & Rohaeti, E. (2024). Computational thinking skill level of senior high school students majoring in natural science. *International Journal of Learning, Teaching and Educational Research, 23*(1), 339–359. https://doi.org/10.26803/ijlter.23.1.17

Hurt, T., Greenwald, E., Allan, S., Cannady, M. A., Krakowski, A., Brodsky, L., Collins, M. A., Montgomery, R., & Dorph, R. (2023). The computational thinking for science (CT-S) framework: Operationalizing CT-S for K-12 science education researchers and educators. *International Journal of STEM Education, 10*(1), Article 1. https://doi.org/10.1186/s40594-022-00391-7

Jin, H.-Y., & Cutumisu, M. (2023). Predicting pre-service teachers' computational thinking skills using machine learning classifiers. *Education and Information Technologies 28*, 11447–11467. https://doi.org/10.1007/s10639-023-11642-7

Kakavas, P., & Ugolini, F. C. (2019). Computational thinking in primary education: A systematic literature review. *Research on Education and Media, 11*(2), 64–94. https://doi.org/10.2478/rem-2019-0023

Kamak, L. P., & Mago, V. (2023). Assessing the impact of using Python to teach computational thinking for remote schools in a blended learning environment. In P. Zaphiris, & A. Ioannou (Eds.), *Learning and collaboration technologies* (vol. 14041, pp. 482–500). Springer. https://doi.org/10.1007/978-3-031-34550-0_35

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika, 54*(4), 681–697. https://doi.org/10.1007/BF02296403

Korkmaz, Ö., Çakir, R., & Özden, M. Y. (2017). A validity and reliability study of the computational thinking scales (CTS). *Computers in Human Behavior, 72*, 558–569. https://doi.org/10.1016/j.chb.2017.01.005

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement, 30*(3), 607–610. https://doi.org/10.1177/001316447003000308

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Li, X., Sang, G., Valcke, M., & Van Braak, J. (2024). The development of an assessment scale for computational thinking competence of in-service primary school teachers. *Journal of Educational Computing Research, 62*(6), 1318–1347. https://doi.org/10.1177/07356331241254575

Li, Y., Schoenfeld, A. H., diSessa, A. A., Graesser, A. C., Benson, L. C., English, L. D., & Duschl, R. A. (2020). Computational thinking is more about thinking than computing. *Journal for STEM Education Research, 3*(1), 1–18. https://doi.org/10.1007/s41979-020-00030-2

Mohaghegh, M., & McCauley, M. (2016). Computational thinking: The skill set of the 21st century. *International Journal of Computer Science and Information Technologies, 7*(3), 1524–1530.

Moreno-Leon, J., Roman-Gonzalez, M., & Robles, G. (2018). On computational thinking as a universal skill: A review of the latest research on this ability. In *Proceedings of the 2018 IEEE Global Engineering Education Conference* (pp. 1684–1689). IEEE. https://doi.org/10.1109/EDUCON.2018.8363437

Niousha, R., Saito, D., Washizaki, H., & Fukazawa, Y. (2023). Investigating the effect of binary gender preferences on computational thinking skills. *Education Sciences, 13*(5), Article 433. https://doi.org/10.3390/educsci13050433

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics, 51*(9), 1352–1375. https://doi.org/10.1080/00140130802170387

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Ocampo, L. M., Corrales-Álvarez, M., Cardona-Torres, S. A., & Zapata-Cáceres, M. (2024). Systematic review of instruments to assess computational thinking in early years of schooling. *Education Sciences, 14*(10), Article 1124. https://doi.org/10.3390/educsci14101124

Ogegbo, A. A., & Ramnarain, U. (2022). A systematic review of computational thinking in science classrooms. *Studies in Science Education, 58*(2), 203–230. https://doi.org/10.1080/03057267.2021.1963580

Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. Basic Books.

Popat, S., & Starkey, L. (2019). Learning to code or coding to learn? A systematic review. *Computers & Education, 128*, 365–376. https://doi.org/10.1016/j.compedu.2018.10.005

Putra, Z. H., Ramiati, Zufriady, Hidayat, R., Jismulatif, Hermita, N., & Sulistiyo, U. (2022). Development of computational thinking tasks based on Riau Malay culture: A study of fifth-grade public school students in Pekanbaru, Indonesia. *Education 3–13, 52*(8), 1387–1397. https://doi.org/10.1080/03004279.2022.2150063

Rahimi, A. R., & Sevilla-Pavón, A. (2025). Scaling up computational thinking skills in computer-assisted language learning (CTsCALL) and its fitness with language learners' intentions to use virtual exchange: A bi-symmetric approach. *Computers in Human Behavior Reports, 17*, Article 100607. https://doi.org/10.1016/j.chbr.2025.100607

Revana, G., Kavita, K., & Madhavi, V. (2021). Exploring the concept of computational thinking in STEM education. In *Proceedings of the 2021 World Engineering Education Forum/Global Engineering Deans Council* (pp. 375–380). https://doi.org/10.1109/WEEF/GEDC53299.2021.9657230

Romero, M., Lepage, A., & Lille, B. (2017). Computational thinking development through creative programming in higher education. *International Journal of Educational Technology in Higher Education, 14*(1), Article 42. https://doi.org/10.1186/s41239-017-0080-z

Rottenhofer, M., Kuka, L., Leitner, S., & Sabitzer, S. (2022). Using computational thinking to facilitate language learning: A survey of students' strategy use in Austrian secondary schools. *IAFOR Journal of Education, 10*(2), 51–70. https://doi.org/10.22492/ije.10.2.03

Saidin, N. D., Khalid, F., Martin, R., Kuppusamy, Y., & Munusamy, N. A. (2021). Benefits and challenges of applying computational thinking in education. *International Journal of Information and Education Technology, 11*(5), 248–254. https://doi.org/10.18178/ijiet.2021.11.5.1519

Sari, U., Ulusoy, A., & Pektaş, H. M. (2025). Computational thinking in science laboratories based on the flipped classroom model: Computational thinking, laboratory entrepreneurial and attitude. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-024-10192-y

Snyder, S., & Sheehan, R. (1992). The Rasch measurement model: An introduction. *Journal of Early Intervention, 16*(1), 87–95. https://doi.org/10.1177/105381519201600108

So, H.-J., Jong, M. S.-Y., & Liu, C.-C. (2020). Computational thinking education in the Asian Pacific Region. *The Asia-Pacific Education Researcher, 29*(1), 1–8. https://doi.org/10.1007/s40299-019-00494-w

Sovey, S., Osman, K., & Mohd Matore, M. E. E. (2022). Rasch analysis for disposition levels of computational thinking instrument among secondary school students. *Eurasia Journal of Mathematics, Science and Technology Education, 18*(3), Article em2088. https://doi.org/10.29333/ejmste/11794

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan* [Application of Rasch modelling to educational assessment]. Trim Komunikata Publishing House.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Tabesh, Y. (2017). Computational thinking: A 21st century skill. *Olympiads in Informatics, 11*(2), 65–70. https://doi.org/10.15388/ioi.2017.special.10

Taherdoost, H. (2017). Determining sample size: How to calculate survey sample size. *International Journal of Economics and Management Systems, 2*, 237–239.

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

Tariq, R., Aponte Babines, B. M., Ramirez, J., Alvarez-Icaza, I., & Naseer, F. (2025). Computational thinking in STEM education: Current state-of-the-art and future research directions. *Frontiers in Computer Science, 6*. https://doi.org/10.3389/fcomp.2024.1480404

Tripon, C. (2022). Supporting future teachers to promote computational thinking skills in teaching STEM–A case study. *Sustainability, 14*(19), Article 12663. https://doi.org/10.3390/su141912663

Tsai, M.-J., Liang, J.-C., & Hsu, C.-Y. (2021). The computational thinking scale for computer literacy education. *Journal of Educational Computing Research, 59*(4), 579–602. https://doi.org/10.1177/0735633120972356

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

Voskoglou, M. G., & Buckley, S. (2012). Problem solving and computational thinking in a learning environment. *Egyptian Computer Science Journal, 36*(4), 28–46.

Waterman, K. P., Goldsmith, L., & Pasquale, M. (2020). Integrating computational thinking into elementary science curriculum: An examination of activities that support students' computational thinking in the service of disciplinary learning. *Journal of Science Education and Technology, 29*(1), 53–64. https://doi.org/10.1007/s10956-019-09801-y

Weintrop, D., Morehouse, S., & Subramaniam, M. (2021). Assessing computational thinking in libraries. *Computer Science Education, 31*(2), 290–311. https://doi.org/10.1080/08993408.2021.1874229

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

Wu, M., & Adams, R. J. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Educational Measurement Solutions.

Wu, T.-T., Asmara, A., Huang, Y.-M., & Permata Hapsari, I. (2024). Identification of problem-solving techniques in computational thinking studies: Systematic literature review. *SAGE Open, 14*(2). https://doi.org/10.1177/21582440241249897

Xu, Z., & Zhang, J. (2021). *Computational thinking: A perspective on computer science* (1st ed.). Springer. https://doi.org/10.1007/978-981-16-3848-0

Yamashita, T. (2022). Analyzing Likert scale surveys with Rasch models. *Research Methods in Applied Linguistics, 1*(3), Article 100022. https://doi.org/10.1016/j.rmal.2022.100022

Yang, D., Snelson, C., & Feng, S. (2023). Identifying computational thinking in students through project-based problem-solving activities. *Information Discovery and Delivery, 51*(3), 293–305. https://doi.org/10.1108/IDD-09-2022-0091

Yang, T.-C., & Lin, Z.-S. (2024). Enhancing elementary school students' computational thinking and programming learning with graphic organizers. *Computers & Education, 209*, Article 104962. https://doi.org/10.1016/j.compedu.2023.104962

Zamzami, E. M., Tarigan, J. T., Zendrato, N., Muis, A., Yoga, A. P., & Faisal, M. (2020). Exercising the students computational thinking ability using Bebras challenge. *Journal of Physics: Conference Series, 1566*(1), Article 012113. https://doi.org/10.1088/1742-6596/1566/1/012113

Zapata, J. H., Gutiérrez Posada, J. E., & Diago, P. D. (2024). Design and validation of a computational thinking test for children in the first grades of elementary education. *Multimodal Technologies and Interaction, 8*(5), Article 39. https://doi.org/10.3390/mti8050039