Heliyon 12 (2026) e45423

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

# Characteristics and psychometric properties of computational thinking assessments in children and adolescents: A systematic review

Carolina Robledo-Castro [a,b,*], Gabriele Pozzan [c], Chiara Montuori [c], Tullio Vardanega [c], Barbara Arfè [c]

[a] Universidad del Tolima, Colombia
[b] Universidad Autónoma de Manizales, Colombia
[c] University of Padova, Italy

ARTICLE INFO

*Keywords:*
Computational thinking
Assessments
Validity
Reliability
Systematic review

ABSTRACT

*Background:* Computational thinking is a problem-solving approach that utilizes computer science principles, enabling a computer to carry out the solution. Over the past decade, computational thinking has become essential in 21st-century education, particularly in K-12 curricula. Assessment tools have been growing in numbers for computational thinking in K-12 education, often equating it with programming or digital skills. However, these assessments frequently lack consensus regarding what should be measured and often fail to meet standard psychometric criteria. To address this, the present study conducted a systematic review of the assessment instruments available to date, analyzing their characteristics, psychometric qualities, and underlying theoretical constructs with the aim of identifying instruments that demonstrate robustness and reliability, as well as highlighting new lines of research in this area.

*Method:* This systematic review followed the parameters of the PRISMA statement guidelines. The reviewed studies were identified and retrieved from the following electronic databases and publication platforms: Science Direct, Springer Link, Taylor & Francis, Sage, IEEE Xplore, Web of Science, and Psychinfo. The automatic filters applied included research articles and full-text publications. The review protocol was prospectively registered in INPLASY202340069.

*Results:* Sixty-four studies met the eligibility criteria and reported on 65 assessment instruments. Most instruments were questionnaires and task-based rubrics, with elementary school children being the most frequently targeted population. Of the included studies, 48 (75%) reported at least one measure of reliability or validity; 36 (56%) reported reliability evidence and 36 (56%) reported validity evidence. At the study level, internal consistency was the most frequently reported reliability measure (42.2%), followed by inter-rater reliability (14.1%), test–retest reliability (6.3%), alternate-form reliability (4.7%), split-half reliability (1.6%), and plausible values reliability (3.1%). Criterion validity (26.6%) and content validity (23.4%) were the most frequently reported forms of validity evidence, followed by construct validity (21.9%) and face validity (9.4%). The instruments with the strongest psychometric evidence were further examined, while those requiring additional validation were identified.

*Conclusion:* The review identifies numerous tools for assessing computational thinking in K-12 education, showing a wide variety in the computational thinking components they aim to

* Corresponding author. Universidad del Tolima, Colombia
*E-mail address:* crobledoc@ut.edu.co (C. Robledo-Castro).

https://doi.org/10.1016/j.heliyon.2026.e45423
Received 6 June 2024; Received in revised form 5 September 2026; Accepted 7 September 2026
Available online 10 September 2026

evaluate. While some studies report on the reliability of these tools, evidence of their validity is less frequently documented.

## 1. Introduction

According to Wing [1], Computational Thinking (hereafter CT) refers to the ability to solve problems using computational principles. It involves the thought processes required to formulate problems and devise solution strategies that can be successfully executed by a computational agent, whether human or machine [2]. CT mainly involves the following steps: analyzing the problem, reducing its complexity by breaking it down into smaller units, expressing and representing the solution through executable algorithms or sequences of instructions, and finally, debugging – i.e., diagnosing and resolving errors in algorithms, programs, or processes – and verifying the achievement of the goal [2–5].

To date, there is still no full consensus on the exact components of CT, which has made it difficult to establish uniform criteria for designing teaching programs and assessment tools for CT [5–8]. Previous attempts to define the components of CT suggest six main components: decomposition, abstraction, algorithm design, debugging, and generalization or pattern recognition [5].

In their systematic review, Selby and Woollard [7] identified three recurring terms in definitions of CT: the idea of a thought process, the concept of abstraction, and the concept of decomposition. Other frequently mentioned terms included modelling, simulation, and visualization, but these reflected peripheral categories rather than core components of CT [6]. Based on their review, Selby and Woollard [7] proposed a five-component model: decomposition, abstraction, algorithmic thinking, debugging, and generalization.

### 1.1. Computational thinking in K-12 education

The term Computational Thinking has been defined from multiple perspectives over time, evolving according to its applicability and purpose. To facilitate its analysis and interpretation, some authors have categorized CT definitions into three types: generic, operational, and educational-curricular [4].

Generic definitions refer to the classical conceptualization proposed by Wing [1,2], who describes CT as the ability to solve problems and design solutions using computational principles. In this view, CT encompasses thought processes involved in formulating and solving problems in such a way that solutions can be represented as computational steps and algorithms and carried out by either a human or a computational agent [2,3,8].

Operational definitions of CT aim to standardize its conceptualization and application by providing an explanatory framework that defines its scope. An example of this is the definition proposed by the Computer Science Teachers Association -CSTA- [9] and the International Society for Technology in Education -ISTE-, which conceptualize CT as a problem-solving approach that involves the logical organization of data, abstraction, automation of solutions, and generalization of problem-solving strategies.

Educational-curricular definitions of CT [4] focus on its integration into education by identifying key concepts, pedagogical approaches, and learning dimensions. A representative example of this approach has been proposed by Brennan and Resnick [10] who developed a conceptual framework for teaching and assessing CT, addressing what aspects should be taught and how they should be evaluated. This framework is organized into three dimensions: (1) computational concepts: essential elements to most programming languages used to formulate computational solutions (e.g., example, sequences, loops, parallelism, conditionals, operators, and data); (2) computational practices: the ways in which a programmer interacts with concepts; and (3) computational perspectives: the perceptions that learners develop about the world and themselves when programming. Nonetheless, one can observe that the framework proposed by the authors restricts the teaching of CT to block-based programming languages – specifically, Scratch – thereby limiting its application to other CT learning contexts.

### 1.2. Assessment of computational thinking in K-12 education

Computational Thinking is widely recognized as a critical 21st century skill, particularly for the development of competencies in science, technology, engineering, and mathematics (STEM) skills [5,11,12]. As a result, many countries are prioritizing the teaching of CT skills in K-12 education. The aim is not only for children to become proficient users of technology, but also to learn essential skills such as algorithm design, problem decomposition and modelling. These skills enable students to approach new challenges and develop innovative solutions using the resources of computational systems [5,13,14].

As the importance of CT education continues to grow, researchers have undertaken numerous projects aimed at designing effective learning opportunities, curricula, teaching environments and assessment tools for CT education [15]. However, this endeavor involves three major challenges that require further discussion: (1) determining which aspects of CT should be taught, (2) identifying the most effective learning methods for CT, and (3) developing reliable methods for assessing CT learning.

As the demand for CT education in schools continues to grow, the need for effective and valid assessment tools becomes increasingly evident [15]. These tools not only help identify students' strengths and areas for improvement but also provide essential insights for designing more effective and equitable pedagogical strategies. Moreover, they enable educators to track student progress, evaluate the achievement of learning objectives, and adapt instruction to address specific challenges [16,17]. Beyond the classroom, CT assessment instruments are an essential supporting tool for researchers, policymakers, and practitioners by providing valuable data

to refine educational practices and inform decision making in this evolving field.

In education, CT assessment plays a fundamental role in supporting teaching and learning processes across all educational levels. Structured assessment of CT offers multiple benefits. First, it generates valuable feedback for emerging educational programmes, enabling educators to refine instructional strategies and optimize curriculum design [8]. Second, it supports the effective monitoring of both individual and collective student progress in CT-related skills, thereby facilitating targeted pedagogical interventions [14]. Finally, robust CT-specific assessment instruments enable comparison across intervention programmes and outcomes, and help identify and mitigate digital-skills gaps—thus promoting more equitable learning opportunities for all students [18].

Moreover, a strong theoretical framework to guide the evaluation of CT is crucial to achieving such a goal. This calls for clarifying the dimensions that define CT and distinguishes them from other cognitive skills. According to Guggemos et al. [19], evaluating complex skill sets such as CT requires considering both the structural and conceptual dimensions, allowing for the definition of operationalized learning objectives. Establishing proficiency levels is essential for effectively interpreting and communicating test results; however, further research is needed to define precise levels of CT competence. Despite the growing interest in CT assessment, many different opportunities still remain for research and development in this field [17].

### *1.3. Background on computational thinking assessment*

Several literature reviews have contributed to the ongoing discussion on CT assessment [15,17,20–23]. Lockwood and Mooney [24] examined the inclusion of CT in curricula and the available assessment methods. Their findings revealed that most existing assessment approaches, which primarily focus on problem solving and analytical thinking, remain in the early stages of development, underscoring the need for more comprehensive evaluation frameworks. Other reviews have examined empirical studies on CT assessment instruments [20,22,23], consistently reporting insufficient evidence regarding the reliability and validity of these tools.

Grover et al. [17] identified two primary approaches to assessing CT skills: (1) questionnaires with single-answer items and (2) programming tasks or projects evaluated by teachers. While questionnaires are easier to administer and analyze, they often fail to capture complex problem-solving abilities. In contrast, programming tasks provide deeper insights into students' CT skills but tend to be subjective and time-consuming for educators. Additionally, the study highlighted a lack of assessment tools specifically designed for evaluating CT within block-based programming environments. The authors emphasized the importance of developing objective evaluation instruments to assess key CT skills, such as debugging, code tracking, decomposition, and pattern generalization.

Building on these insights, Cruz Alvez et al. [21] conducted a systematic mapping of assessment approaches that focus on analyzing students' codes. Their research suggested that automated evaluation tools could support educators in assessment and grading while also guiding students in their learning process. However, there is no clear consensus on the most effective assessment criteria or the role of pedagogical feedback.

Tang [23] further categorized CT assessment studies into two domains: (1) first-order constructs, which directly evaluate CT components, and (2) second-order constructs, which assess related fields such as programming, STEM skills, or non-STEM cognitive abilities. Among first-order constructs, 41.7% of the reviewed instruments assessed CT skills, such as Bebras tasks [25], while 7.5% measured perceptions of CT using instruments like those developed by Leonard et al. [26] and Korkmaz et al. [27]. For second-order constructs, many assessment tools focused on programming skills, such as those evaluating Scratch projects [17]. Additionally, some tools assessed related STEM skills, such as the Lattice Land software, which was designed to foster mathematical thinking alongside computational thinking [28], while others explored applications of CT beyond STEM disciplines, including its relationship with poetic thinking in English education [29].

The most frequently assessed CT skills include algorithmic thinking, abstraction, problem decomposition, logical reasoning, and data analysis [15,22]. Among European empirical studies, the most representative assessment tools include task- or challenge-based approaches such as Dr. Scratch [26,30] and Bebras Tasks [25], self-report Likert-scale questionnaires like the CT Ability Scale [27], multiple-choice tests such as the CT Test [4], and coding tasks within visual block-based programming environments such as Scratch, Alice, and Code.org [20]. However, a key limitation of previous systematic reviews is that many of their analyses and conclusions were drawn from studies that did not differentiate among age groups or educational levels [15,22,23]. As noted earlier, another recurring gap in the literature concerns the limited evidence on the validity and reliability of existing CT assessment instruments [15,23]. More recent reviews have documented increasing efforts to incorporate validity evidence into instrument development [18,22,31]. Yet, a closer examination of these studies shows that, although they describe the validation procedures employed and the dimensions of validity and reliability examined [18,23,31], they rarely provide specific results for each instrument or indicate whether the reported evidence was adequate or satisfactory.

In this context, the present systematic review aims to go beyond previous work by comprehensively collecting and analyzing the available empirical data on the psychometric properties of each CT assessment instrument in K-12 education. Through a systematic analysis of the literature, this review not only documents the reported validation but also critically examines the methodological soundness of the instruments, the quality of the evidence presented, and the extent to which validity and reliability criteria have been met. This paper seeks to provide a more complete and detailed overview of the current landscape, offering key information to guide future research and improve the selection of assessment tools in the educational field.

Compared to previous systematic and scoping reviews, in which both self-report instruments aimed at assessing attitudes and perceptions and performance instruments were considered [18,23,31], the present review focused exclusively on instruments designed to assess performance on CT tasks. This decision was made to facilitate comparisons between instruments and constructs.

Overall, the systematic review provides a detailed analysis of existing CT assessment tools, based on students' age level and other characteristics, focusing on the psychometric properties of these performance-based instruments to support conclusions about CT

assessment. This systematic review aimed to: (1) identify and characterize existing instruments designed to assess computational thinking performance in K-12 populations; (2) synthesize the psychometric evidence reported for these instruments, including validity (content, criterion, construct, and face validity) and reliability (internal consistency, test-retest, inter-rater, alternate-form, and split-half reliability); and (3) critically evaluate the methodological rigor and completeness of the evidence reported, in order to identify the most robust instruments currently available and highlight gaps warranting further validation research.

## 2. Method

A systematic review was conducted in accordance with the guideline of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses [32]. This guideline contains a 27-item checklist and is designed to promote a systematic and transparent report of the rationale, methodology, and main findings of the studies. The protocol for this review can be retrieved online at: https://inplasy.com/ under the following reference number INPLASY202340069, https://inplasy.com/inplasy-2023-4-0069/. Although the INPLASY registration was completed in April 2023, the initial database search was carried out in June 2022, prior to the formal registration. To ensure that the review reflected the most current evidence, the search strategy was subsequently re-executed and updated to include studies published up to April 2023. The INPLASY record was updated accordingly to clarify the search timeline and maintain consistency between the preregistered protocol and the procedures reported in this manuscript.

### 2.1. Eligibility criteria

This review applied the PICO guideline (Population, Intervention, Comparison, Outcome) [33] to specify all the eligibility criteria for a study to be included (see Table 1).

Based on these guidelines, the inclusion criteria comprised empirical studies in which CT was evaluated, as well as validation studies that presented and tested new CT assessment tools, focusing on K–12 student populations. Studies published between 2006 and 2023 were considered eligible, reflecting the period following Wing's [1] seminal introduction of CT into educational research. The exclusion criteria included abstracts, conference proceedings, review papers, and theoretical articles, as well as studies assessing constructs other than CT knowledge or skills—such as self-efficacy, attitudes, or perceptions. Additionally, studies evaluating teachers rather than students were excluded.

### 2.2. Information sources and search

The primary sources for the search were major electronic databases and publishing platforms, including ScienceDirect, Springer Link, Taylor & Francis, Sage, and IEEE Xplore. In addition, PsycINFO and Web of Science were incorporated into the search strategy, even though they were not listed in the original INPLASY protocol, as both databases provide complementary coverage of educational, cognitive, and psychological research relevant to computational thinking assessment. These databases were selected based on the relevance and quality of their indexed publications, as well as their comprehensive coverage of the study domains. The search strategy employed the descriptors "Computational Thinking," "Coding learning," "Assessment," "Measurement," and "Test," which were combined using Boolean operators "AND" and "OR." The resulting search string was: ("Computational thinking" OR "coding learning") AND ("assessment" OR "measurement" OR "test"), with minor adjustments applied to accommodate the specific syntax requirements of each database. To enhance completeness and reduce the risk of missing eligible studies, the reference lists of all included articles were also hand-searched.

### 2.3. Selection process

The total search results from each database were transferred to the Mendeley® reference management software, where duplicates were removed to ensure that each study was considered only once. In the selection phase, the first author performed an initial filter based on the review of titles and abstracts, with the aim of excluding those studies that clearly did not meet the previously established inclusion criteria. Subsequently, a second reviewer independently assessed the list of preselected articles to verify their relevance and confirm their eligibility. During this process, the second reviewer approved the inclusion of 95.4% of the articles suggested by the first reviewer, showing an inter-rater Kappa coefficient of $\kappa = .812$, suggesting a high level of inter-rater agreement. Subsequently, the first author proceeded to perform a reading of each complete article to confirm whether the articles met two specific conditions: 1) the inclusion criteria and 2) proximity to the research question. The articles that met these conditions were reviewed and discussed further in plenary sessions, where two thirds of the authors participated. As a result, 79.1% of the total of the articles found in the preliminary review were approved to proceed with the data extraction.

**Table 1**
PICO guidelines.

| | |
|---|---|
| Participants (P) | The participants of the studies had to be in K-12 education range. |
| Intervention (I) | The purpose of the review was not interventions but tools to measure computational thinking. |
| Comparison (C) | Does not apply |
| Outcome (O) | Data on validity and reliability of the computational thinking assessment tools. |

### 2.4. Data extraction

To systematize the extracted information, a documentary matrix was constructed capturing the following data from each article: author, title, year, country, sample characteristics, objectives, instrument features, CT dimensions assessed, and validity and reliability evidence. Full methodological details of the review protocol are reported in Robledo-Castro et al. [34].

### 2.5. Quality assessment

In this systematic review, quality was assessed at the instrument rather than at the study level. This choice was made because the review does not evaluate study findings. Instead, it examines the psychometric quality of the CT assessment tools used across studies. Consequently, the quality of the reported information (assessment tools characteristics) is independent of the methodological quality or potential design flaws of the individual studies from which they were drawn [35].

To judge the quality of each CT assessment tool, we examined the completeness, clarity, and adequacy of the information reported regarding its psychometric properties and measurement characteristics. Quality judgments were guided by widely used standards for evaluating measurement instruments [36].

Both qualitative and quantitative criteria were applied.

- Content validity: Considered satisfactory when the authors provided a clear conceptual definition of CT, and showed evidence that items/tools adequately represented the intended CT dimensions.
- Construct and criterion validity: Judged as satisfactory when appropriate statistical methods were used (e.g., exploratory or confirmatory factor analysis) and/or when the measure shows convergence with other measures of the same construct
- Internal consistency: Considered acceptable when reliability coefficients (e.g., Cronbach's α) for each subscale were $\geq .70$, assuming unidimensionality of the construct.
- Other reliability evidence (e.g., test–retest, inter-rater): Judged satisfactory when appropriate indices were reported (e.g., ICC $\geq .70$ for stability over time or scorer agreement).

In addition, to provide a comprehensive description of the study sources from which we retrieved the above mentioned quality indicators, we also considered the following.

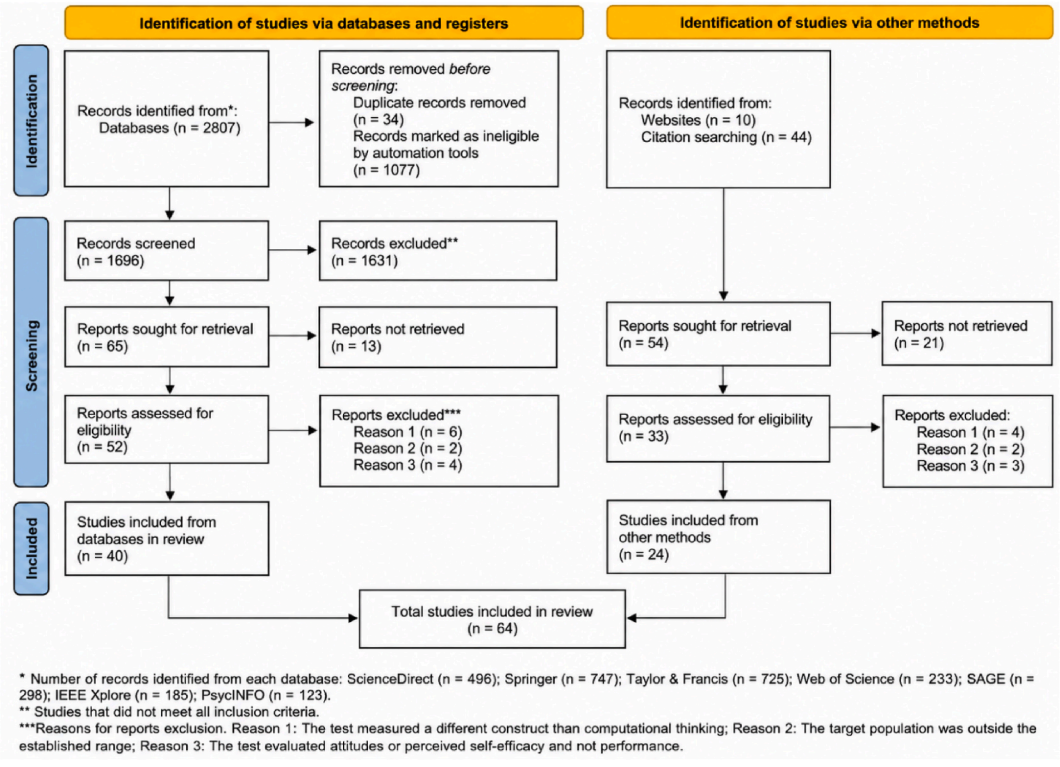


**Fig. 1.** Flow chart of the article selection procedure with references to PRISMA 2020 guidelines.

- Sample description: Evaluated based on the adequacy and clarity of reporting about the population on which the instrument was tested (e.g., age, educational level, sample size). Psychometric results were considered stronger when based on sufficiently large and well-characterized samples.
- Measurement procedures: Judged satisfactory when the administration mode, scoring method, and task requirements were described in enough detail.

Based on these criteria, each instrument was classified according to the level of evidence available for its psychometric soundness. A summary of the analyzed instruments' quality assessment is reported at the end of the results section (Fig. 5).

## 3. Results

A structured search of the selected academic databases (see Section 2.2 for the full search strategy) initially yielded 2807 records (see Fig. 1). After removing 34 duplicate records and excluding 1077 records through automated screening procedures (e.g., language and document-type filters), 1696 records remained for title and abstract screening. Following this screening, 1631 records were excluded, leaving 65 reports sought for retrieval. Of these, 13 reports could not be retrieved (e.g., because the full text was unavailable or could not be accessed through institutional subscriptions), resulting in 52 reports assessed for eligibility through full-text review.

Full-text evaluation was conducted by the research team and subsequently discussed in plenary meetings with all co-authors. At this stage, 12 reports were excluded for predefined reasons (see Section 2.3), including measuring constructs other than computational thinking, involving populations outside the established age range, or focusing on attitudes and perceived self-efficacy rather than performance-based measures. As a result, 40 studies identified through database searches were included in the review.

As specified in the preregistered protocol, the review process continued beyond preregistration to ensure comprehensive coverage of the literature. Accordingly, an additional 54 records were identified through citation searching and other secondary sources. Of these, 33 reports were sought for full-text review, while 21 could not be retrieved. Following eligibility assessment, 9 of the 33 retrieved reports were excluded for predefined reasons, resulting in 24 additional studies meeting the inclusion criteria. In total, 64 studies were included in the final review. Fig. 1 presents the PRISMA 2020 flow diagram illustrating the study selection process [32].

### 3.1. Characteristics of the studies

A summary of the included studies and the instruments evaluated is presented in Table 2. For each study, the table provides information on: (1) the name of the computational thinking (CT) assessment instrument developed or used; (2) the type of instrument; (3) the context or modality in which the instrument was administered, including block-based programming platforms, robotics-based contexts, or unplugged CT activities; (4) the target population and sample size; and (5) the reported evidence of validity and reliability. Examination of the instruments reported across the included studies showed that one study, Sung [37], evaluated and reported psychometric evidence for two distinct instruments. Therefore, although n = 64 studies were included in the review, these studies collectively provided data on n = 65 CT assessment instruments.

Based on the analysis of the studies included in this systematic review, most standardized instruments for assessing computational thinking are designed for primary school students—particularly those in grades 3 to 6—making this the predominant target population in the existing literature. Additionally, the most common type of instrument used in standardized CT assessments consists of questionnaires with multiple-choice, task-based items, reflecting a preference in the literature for structured, easily scorable formats.

The publication dates of the included studies show a clear increase in research building, adapting, and validating CT assessment instruments, particularly from 2020 onward (see Fig. 2). Regarding geographical distribution (Fig. 3), the United States produced the largest number of studies (n = 23), followed by China (n = 9) and Spain (n = 9), Taiwan (n = 5), Turkey (n = 3) and South Korea (n = 3), and Japan, Indonesia, Switzerland, Lithuania, and Estonia (n = 2 each). Cyprus, France, Slovenia, Croatia, and Oman contributed one study each. No studies from Latin America, Africa, or Australia met the inclusion criteria.

Regarding the characteristics of the included studies (N = 64), most targeted primary school students (67.2%, n = 43), followed by secondary school students (34.4%, n = 22) and preschool students (21.9%, n = 14). All included studies assessed students' computational thinking (CT) skills through performance-based measures, as studies focusing exclusively on other assessment dimensions were excluded from the review. Most studies (87.5%, n = 56) assessed CT performance through problem-solving tasks, whereas a smaller proportion (9.4%, n = 6) evaluated performance through game-programming or robotics projects. Because some studies included more than one target population or assessment approach, these categories were not mutually exclusive and their percentages may therefore exceed 100%.

At the instrument level (N = 65), questionnaires comprising problem- or task-based multiple-choice items were the most common assessment format (46.2%, n = 30). Instruments consisting of exercises or tasks requiring a single open-ended response accounted for 21.5% (n = 14), while observation rubrics completed by teachers or researchers to assess students' performance on individual tasks or projects represented 32.3% (n = 21). In addition, 9.2% (n = 6) of the instruments consisted of software-based assessment tools that automatically recorded and evaluated students' performance on programming tasks within the application. Regarding the activity modality associated with the instruments, 56.9% (n = 37) involved plugged activities, primarily block-based programming, 53.8% (n = 35) involved unplugged activities, and 7.7% (n = 5) incorporated both plugged and unplugged approaches. Because individual instruments could involve more than one format or activity modality, these categories were not mutually exclusive.

**Table 2**
Characteristics of the studies included in the systematic review.

| Author (Year), Country | Instrument name (initials) | Type of Instruments, Type of items (Platform/device) | Population, Sample size[a] | Validity | Reliability |
|---|---|---|---|---|---|
| Angeli & Valanides [38], Cyprus | Rubric for assessing children's CT in a holistic way | Rubric, task-based (Bee-Bot) | Preschool n = 50 | NR | Inter-rater |
| Basu et al. [39], USA | NR | Task-based | 6th grade n = 25 | NR | NR |
| Basu et al. [40], USA | CT Practices and concepts assessment instrument | Questionnaire, Task-based with multiple choice response (CoolThink@JC) | From 4th to 6th grade n = 1808 | Content | Internal consistency |
| Bers [41], USA | NR | Rubric, Task-based with Likert scale (TangibleK robots) | Preschool n = NR | NR | NR |
| Bubica & Boljat [42], Croatia | Croatian CT assessment | Questionnaire and Rubric, task-based with multiple choice response, essay and short answer (Unplugged) | From 4th to 6th grade n = 352 | Criterion | Internal consistency |
| Chen et al. [13], USA | Evaluation of CT of primary school students in everyday reasoning and robotic programming | Rubric and Questionnaire, Task-based with multiple-choices and open-ended questions (Plugged, Robotic) | 5th grade n = 121 | | Internal consistency, Alternative forms |
| Choi et al. [43], Korea | Lee (1999) adaptation | Questionnaire and rubric, Task-based Multiple-choices and open-ended questions (unplugged, puzzles) | From 4th to 6th grade n = 82 | Content | |
| Clarke-Midura et al. [44], USA | Kindergarten CT assessment | Rubric, Task-based (unplugged, Coding in Kindergarten "CiK project") | Preschool n = 89 | | Internal consistency, Inter-rater |
| Clarke-Midura et al. [45], USA | KRO indicators | Rubric, Task-based (unplugged) | Preschool n = 17 | | Inter-rater |
| Coban & Korkmaz [46], Turkey | Computational Thinking Score (CTS) | Software, task-based (plugged, emrecoban.com) | From 9th to 12th grade n = 203 | Content, Criterion, Construct | Internal consistency |
| Dagienė et al. [47], Lithuania | Bebras tasks | Task-based (Unplugged) | K-9 Primary school and High school n = NR | NR | NR |
| El-Hamamsy et al. [48], Switzerland & Spain | The competent Computational Thinking Test (cCTt) | Questionnaire, Task-based with multiple choice question (unplugged) | Upper Primary School n = 1519 | Content, Construct, Face | Internal consistency |
| Gane et al. [49], USA | NR | Rubric, Task-based (Plugged, schatch) | From 3rd to 4th grade n = 144 | | Inter-rater |
| Hooshyar et al. [50], Estonia & China | Román-González et al. [4] adaptation | Questionnaire, Task-based with multiple choice question (Plugged, AutoThinking) | 5th grade n = 79 | | |
| Kalyenci et al. [51], Turkey | CodingTest AB - Early Childhood Coding Skills Assessment Test | Rubric, Task-based (Plugged, My school bus robotic kit tool) | Preschool and 1st grade n = 308 | Content, Criterion | Internal consistency |
| Khodabandelou & Alhoqani [52], Oman | CTCs test Román-Gonzalez et al. [4] adaptation | Questionnaire, Project-based with multiple choice question (LEGO robots) | 5th grade n = 35 | Face | Internal consistency |
| Kong & Lai [53], China | CT concept test | Questionnaire, Task-based with multiple choice question (Scratch) | From 5th to 6th grade n = 13670 | Content, Construct | Internal consistency |
| Kong & Wang [54], China | CT practices test | Questionnaire, Task-based with multiple choice question (unplugged) | From 4th to 6th grade n = 13956 | Construct | |
| Kwon et al. [55], USA | NR | Questionnaire, Task-based with multiple choice question (Scratch) | 6th grade n = 200 | | Internal consistency, Alternative forms |
| Lee & Cho [56], South Korea | Computational Thinking Effectiveness | Questionnaire and interview, Task-based (Bebras task) | Preschool n = NR | Content | |
| Leonard et al. [57], USA | NR | Rubric, Project-based (Plugged, LEGO robots) | From 5th to 8th grade n = 76 | | |
| Li & Traynor [58], USA | Computational thinking competency assessment (CTCA) | Questionnaire, Task-based with Think-aloud question and Multiple-choice question (Unplugged) | 7th grade n = 564 | | |
| Li et al. [59], China | CT Assessment for Chinese Elementary Students (CTA-CES) | Questionnaire, Task-based with multiple choice question (Unplugged, real-life problem) | From 3rd to 6th grade n = 228 | Criterion, | Internal consistency, Parallel forms |
| Matere et al. [60], Taiwan | Román-Gonzalez et al. [4] adaptation | Questionnaire, Task-based with multiple-choices (Plugged, Arduino) | From 4th to 6th grade n = 64 | | Internal consistency |
| Metcalf et al. [61], USA | NR | Rubric, Project-based (Unplugged, EcoMOD curriculum) | 3rd grade n = 94 | | Inter-rater |

(*continued on next page*)

**Table 2** (*continued*)

| Author (Year), Country | Instrument name (initials) | Type of Instruments, Type of items (Platform/device) | Population, Sample size[a] | Validity | Reliability |
|---|---|---|---|---|---|
| Moreno-León et al. [30], Spain | Dr.Scratch | Software, Project-based (Plugged, Scratch) | Aged 10 to 14 n = 109 | | |
| Moreno-León et al. [26], Spain | Dr.Scratch | Software, Project-based (Plugged, Scratch) | Primary school and High school n = 450 | Criterion | |
| Munawarah et al. [62] Indonesia | Test for CT skills in the mathematics-based RME (Realistic Mathematics Education) | Questionnaire, Task-based with multiple choice question and essay (Unplugged, Based on mathematical problems) | 8th grade n = 102 | | Internal consistency |
| Oomori et al. [63], Japan | NR | Questionnaire, task-based (Unplugged and plugged) | Primary school n = 74 | | |
| Ortega-Ruipérez et al. [64] Spain | Instrument to measure CT through complex problem solving | Questionnaire, Task-based (Unplugged, PISA) | High school n = 66 | Construct | |
| Ota et al. [65], Japan | Ninja Code Village | Software, task-based (Plugged, Scratch) | NR | | |
| Palts [66], Estonia | Bebras Challenger | Questionnaire, Task-based (Unplugged, Bebras task) | High school n = 649 | Construct | |
| Piatti et al. [67], Switzerland | CT-cube - Cross Array Task (CAT) | Protocol with rubric, task-based (Unplugged) | Aged 3 to 16 n = 109 | | |
| Polat et al. [68], Turkey | CTt Turkish adaptation to Roman-Gonzalez [4] | Questionnaire, Task-based with multiple choice question (Plugged, code.org) | 5th & 6th grade n = 328 | | Internal consistency |
| Putri et al. [69], Indonesia | Mathematical Computational Thinking Skill (MCTS) | Questionnaire, Task-based with multiple choice question (based on mathematical problems) | High school n = NR | Content | |
| Relkin [70], USA | TACTIC-KIBO | Rubric, Task-based (Plugged, KIBO robot) | Preschool and 1st grade n = 15 | Criterion, Face | Inter-rater |
| Relkin & Bers [71], USA | | | | Criterion | |
| Relkin & Bers [72], USA | TechCheck-K | Questionnaire, Task-based with multiple choice question (Unplugged) | Preschool n = 89 | Criterion | |
| Relkin et al. [73], USA | TechCheck | Questionnaire, Task-based (KIBO robot) (Plugged, KIBO robot) | 1st & 2nd grade n = 768 | Criterion, Face | Internal consistency |
| Rodríguez-Martínez et al. [74], Spain | Computational Thinking Test (CTT) [4] | Questionnaire, Task-based with multiple-choices (Plugged, code.org) | 6th grade n = 47 | | |
| Román-González [4], Spain | Computational thinking test (CTt) | Questionnaire, Task-based with multiple-choices (Plugged, code.org) | 5th to 10th grade n = NR | Content | Internal consistency |
| Román-González [75], Spain | Computational thinking test (CTt) | Questionnaire, Task-based with multiple-choices (Plugged, code.org) | 5th to 10th grade n = 1251 | Criterion | |
| Román-González [76], Spain | Computational thinking test (CTt) | Questionnaire, Task-based with multiple-choices (Plugged, code.org) | 7th & 8th grade n = 314 | Criterion | |
| Rowe et al. [77], USA | CT automatic detector in Zoombinis | Software and rubric, task-based (Plugged, Zoombinis) | From 3rd to 8th grade n = 70 | Criterion, | Inter-rater |
| Rowe et al. [78], USA | Interactive assessments of CT (IACT) | Questionnaire and rubric, task-based (Plugged, Zoombinis) | From 3rd to 8th grade n = 3022 | Criterion, Face | Test-retest |
| de Ruiter & Bers [37], USA | Coding Stages Assessment (CSA) | Questionnaire and interview, task-based with open question (Plugged, Scratch Jr.) | From Preschool, to 3rd grade n = 118 | Criterion, Construct | Inter-rater, Split Half Reliability |
| Seiter & Foreman [79], USA | Progression of early computational thinking (PECT) | Rubric, Projects-based (Plugged, Scratch) | From 1st to 6th grade n = NR | | |
| Sigayret et al. [80], France | Mastery of computational concepts and Algorithmic Test | Questionnaire, Task-based with multiple choice question (Plugged, code.org & Scratch) | From 4th to 9th grade n = 449 | Criterion | Internal consistency |
| Sung [81], Korea | TACTIC-KIBO (Korean Adaptation) | Rubric, Task-based (KIBO robot) | Preschool n = 450 | Criterion, Construct, Face | Internal consistency |
| | Bebras Kards (Korean Adaptation) | Questionnaire, task-based (Bebras Kards) | Preschool n = 450 | Criterion, Construct, Face | Internal consistency |
| Ternik et al. [82] | Bebras task | Task-based (Unplugged) | 6th to 9th n = NR | Content | |
| Tran [83], USA | NR | Questionnaire & Interview, Task-based (unplugged) | 3rd grade n = 183 | | Internal consistency |

(*continued on next page*)

**Table 2** (*continued*)

| Author (Year), Country | Instrument name (initials) | Type of Instruments, Type of items (Platform/device) | Population, Sample size[a] | Validity | Reliability |
|---|---|---|---|---|---|
| Tsai et al. [84], Taiwan | Computational Thinking Scale (CTS) | Questionnaire, Task-based with multiple choice question (Unplugged) | 8th & 9th grade n = 388 | Construct | |
| Tsai et al. [85] Taiwan | | | 8th & 9th grade n = 472 | Construct | Internal consistency |
| Tsai et al. [86] Taiwan | Computational Thinking Test for Elementary School Students (CTT-ES) | Questionnaire, Task-based with multiple choice question (Unplugged) | 5th & 6th grade n = 631 | Criterion, Construct | Internal consistency |
| Wang et al. [87], China | Tool Measuring Coding Ability of Young Children (TMCYC) | Rubric, Task-based (card-based game) | Preschool n = 60 | Content, Construct | Inter-rater, Test-retest, Internal consistency |
| Werner et al. [88], USA | Fairy Assessment | Software and Rubric, Task-based (Alice) | Aged 10 to 14 n = 311 | | |
| Wu et al. [89], Taiwan | NR | Questionnaire, Task-based with multiple choice question (Unplugged) | 5th & 6th grade n = 180 | Content | |
| Zapata-Cáceres et al. [90], Spain | The CTt for Beginners (BCTt) | Questionnaire, task-based with multiple-choices (Unplugged) | Aged 5 to 10 n = 299 | Content | Internal consistency, Test-retest |
| Zhan [91], China | Computational Thinking Ability Test | Rubric, Task-based with multiple-choices (Microbit) | Primary school n = 48 | | Internal consistency |
| Zhang et al. [92], China | Computational Thinking Test for Lower Primary (CTt-LP) | Questionnaire, task-based with multiple-choices (Unplugged) | 2nd grade n = 72 | | Internal consistency |
| Zhang & Wong [8], China | Computational Thinking Test for Lower Primary (CTt-LP) | Questionnaire, task-based with multiple-choices (Unplugged) | From 1st to 3rd grade n = 1225 | Criterion, Content | |
| Zhang & Wong [93], China | Computational Thinking Test for Lower Primary (CTt-LP) | Questionnaire, task-based with multiple-choices (Unplugged) | | Construct | Internal consistency, Test-retest |
| Zhao & Shute [94], USA | MCTt [4] | Questionnaire, Task-based with multiple choice questions (Plugged, code blocks). | 8th grade n = 69 | | Internal consistency |
| Zhong et al. [95], China | Three-Dimensional Integrated Assessment (TDIA) | Rubric, task-based (Alice) | 6th grade n = 144 | | |

Note: PISA: Programme for International Student Assessment-problems. NR: Not reported.
[a] sample size of the study, not of the standardization.

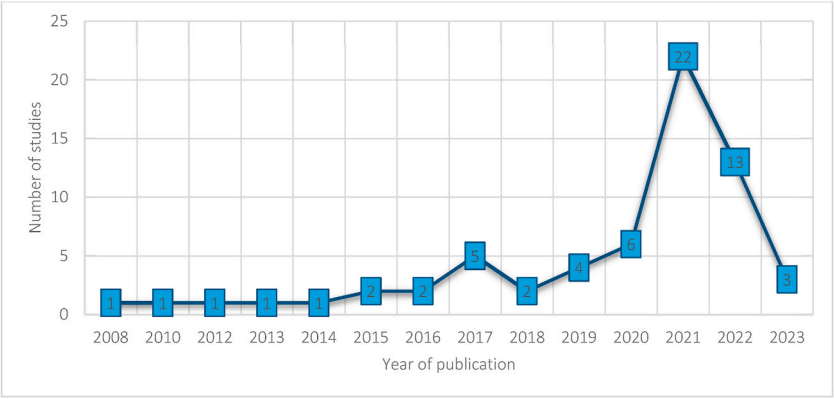


**Fig. 2.** Number of articles published by year.

### 3.2. Conceptual and theoretical bases of the evaluation tool

The CT components assessed varied significantly between studies. This discrepancy may stem from the lack of consensus within the academic community regarding the precise definition of CT components. Nevertheless, most studies included subscales derived from operational definitions of CT, assessing components proposed by Wing [2] and Selby & Woollard [7], such as algorithmic thinking (66%), the capability to develop a step-by-step solution or a set of rules to solve a problem systematically; decomposition (36%), which involves breaking down complex problems into smaller, more manageable parts; abstraction (39%), the process of identifying essential details while ignoring irrelevant information to focus on the core structure of a problem; generalization (28%), the ability to apply learned concepts and solutions to new, similar problems; and debugging (42%), the process of identifying, analyzing, and correcting

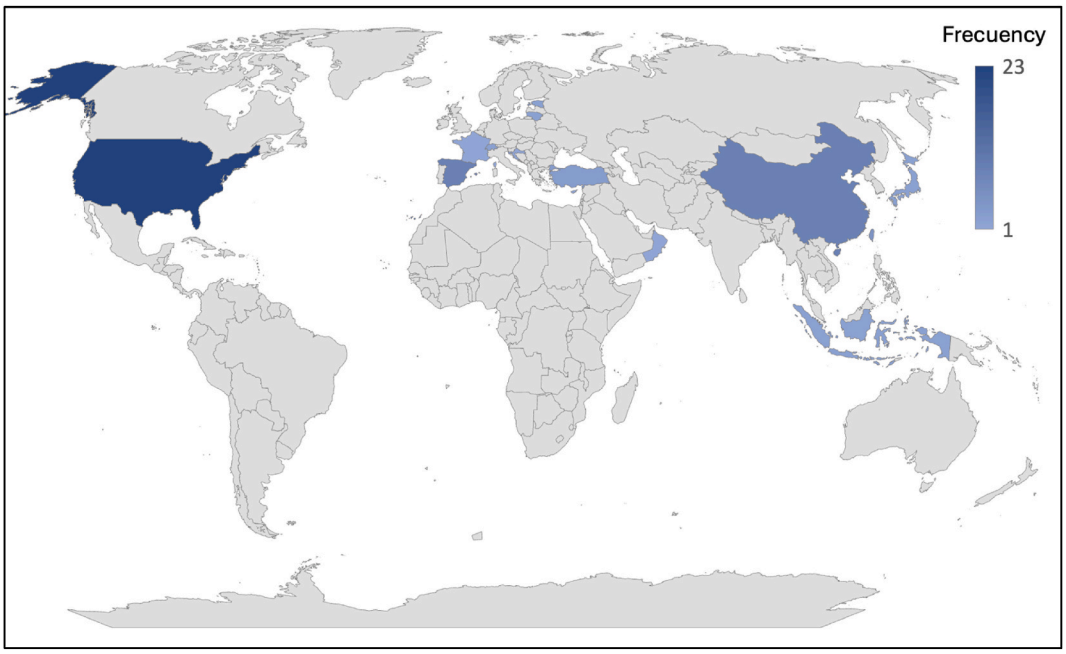


**Fig. 3.** Geographical distribution of studies
Note: The color intensity represents the number of studies conducted in each country. Darker shades indicate a higher number of studies.

errors in a solution or computational process.

Furthermore, other studies reported instruments based on educational-curricular definitions, such as those proposed by Brennan & Resnick [10] (Fig. 4). These definitions focus on algorithm design and encompass key concepts: sequences (36%), which involve executing a series of steps or instructions; loops (36%), which repeat sequences multiple times; events (14%), which trigger specific actions; conditionals (34%), which enable decision-making based on given conditions, allowing for multiple outcomes; parallelization (8%), which allows the simultaneous execution of instruction sequences; functions (12%), independent program segments that perform specific operations; and variables (9%), data elements whose values may change during program execution.

Some studies included less frequent components that, although related to CT, are considered peripheral concepts that complement the core concepts and facilitate their application in various contexts [6]. These components included iteration (5%), which involves repeating design processes to refine solutions until an optimal result is achieved [5]; modularization (14%), defined as constructing larger structures by assembling smaller components [10]; data representation (11%), referring to methods used to express information within a computer system; and simulation and modeling (5%), which involve creating abstract representations to simulate a system's behavior [5]. Additionally, a significant portion of the studies (17%) focused on assessing problem-solving skills.

Fig. 4 presents the factors included in the evaluated instruments, categorized by the type of definitions they are based on, along with their frequency across the studies.

### *3.3. Validity and reliability tests*

Of the final 64 studies, 75% (n = 48) reported at least one measure of reliability or validity. Specifically, 56% (n = 36) reported reliability measures and 56% (n = 36) validity measures. The evidence reported by each study is presented in detail below.

#### *3.3.1. Reliability*

According to classical test theory, reliability refers to the degree to which test scores are consistent, stable, and free from measurement error, representing the proportion of true score variance relative to observed variance [96]. Reliability is operationalized through statistical coefficients that quantify different forms of score consistency [36]. In this review, reliability evidence reported by the included studies aligns with three widely recognized coefficient types: (1) internal consistency, which examines the coherence among test items or subscales; (2) test–retest reliability, which evaluates the stability of scores over time; and (3) inter-rater reliability, which assesses the level of agreement between independent raters. Additional forms such as alternative-form and split-half reliability are also considered in some instruments to evaluate equivalence and internal homogeneity [31,97].

On the other hand, other reliability measures emphasize different types of measurement precision and consistency. According to Item Response Theory (IRT), the accuracy with which a test measures a latent variable depends on the degree to which it captures the examinee's skill level. Rather than relying on a single reliability coefficient, IRT evaluates the Test Information Function (TIF), which indicates how much information a set of items provides across different levels of ability, with greater information corresponding to

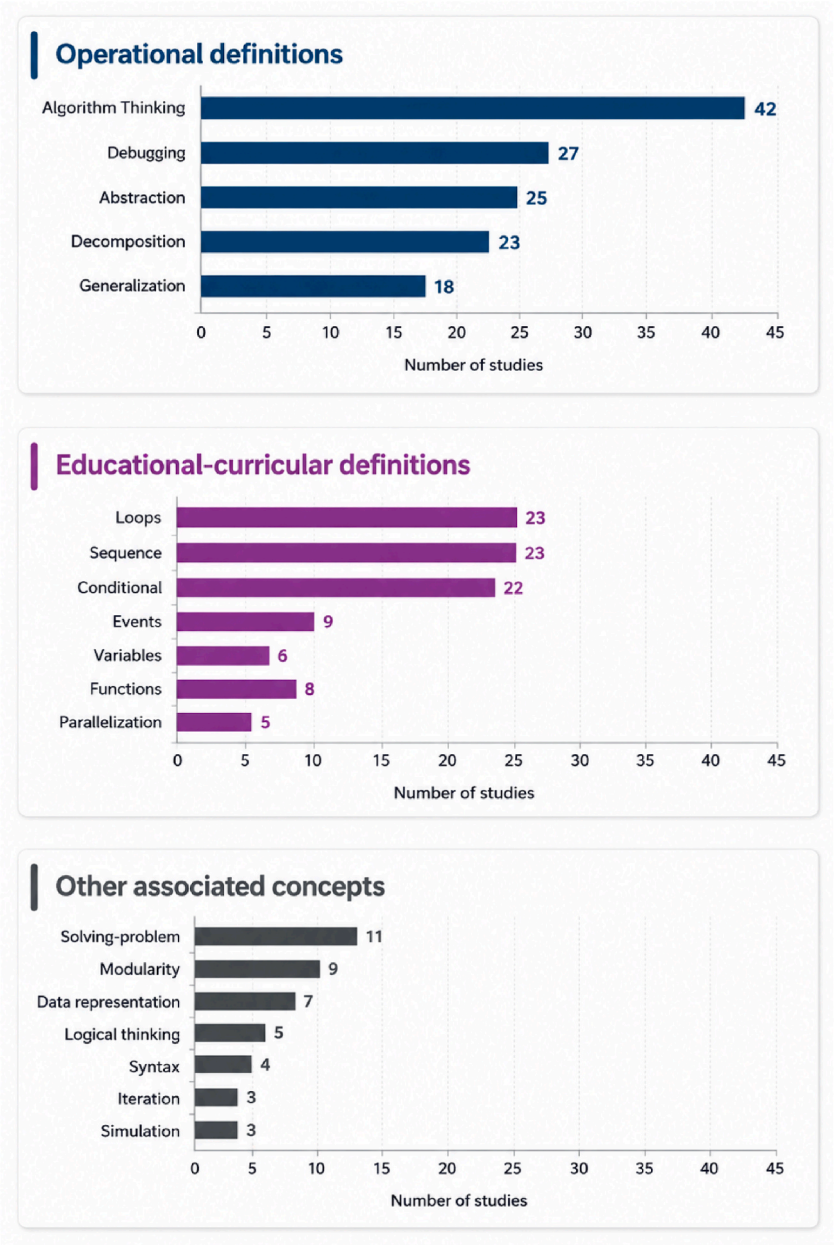


**Fig. 4.** Computational thinking factors assessed.

lower measurement error and higher precision [98]. One of the estimation methods commonly used in IRT is the Weighted Likelihood Estimate (WLE), which assesses an individual's ability level while reducing estimation bias and supporting the differentiation of examinees across varying skill levels [99]. Another frequently used estimator is the Expected A Posteriori/Plausible Values (EAP/PV), which reflects the stability and precision of the latent trait estimation obtained through the measurement instrument [98].

Among the studies reviewed, internal consistency was the most frequently reported reliability measure (42.2%, n = 27), followed by inter-rater reliability (14.1%, n = 9), test-retest reliability (6.3%, n = 4), alternate-form reliability (4.7%, n = 3), split-half reliability (1.6%, n = 1), and reliability of plausible values (3.1%, n = 2).

*3.3.1.1. Internal consistency.* Internal consistency (IC) is not an inherent property of a scale but rather reflects the consistency of instrument scores within a specific target population; therefore, it may vary across populations and contexts [100]. It is typically assessed from a single administration of an instrument and reflects the degree of consistency among its items. Statistical coefficients commonly used to assess internal consistency include Cronbach's alpha (α), the Kuder–Richardson formula (KR-20), reliability of plausible values (EAP/PV), and McDonald's omega (ω). In general, coefficients of .70 or higher are considered acceptable, although

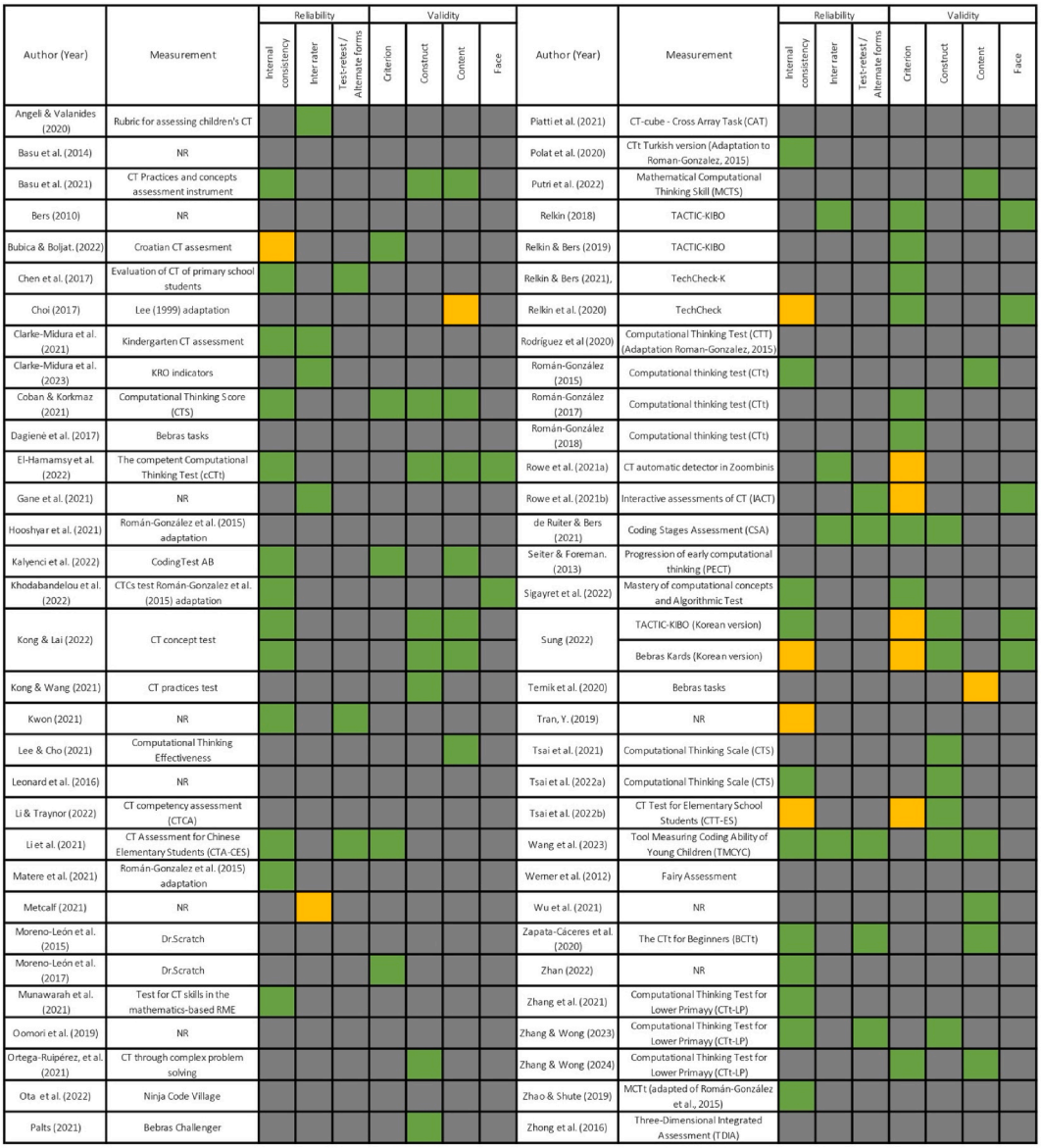


**Fig. 5.** Level of evidence for validity and reliability.

values as low as .65 may be considered acceptable under certain circumstances [101].

Among the 64 included studies, 27 (42.2%) reported evidence of internal consistency, corresponding to 28 of the 65 instruments reviewed (43.1%). Cronbach's alpha was the most frequently reported measure (24 studies, 37.5%) [40,42,46,48,52,53,55,59,60,62,68,73,75,76,80,81,83,86,87,90–94]. Two studies used the Kuder–Richardson formula [44,51], while two studies reported EAP/PV reliability [13,86]; one of these [86] reported both Cronbach's alpha and EAP/PV reliability.

Of the 28 instruments for which internal consistency evidence was reported, 21 (32.3%) showed acceptable coefficients, with values ranging from .70 to .98. In contrast, five instruments reported Cronbach's alpha coefficients ranging from .48 to .68 [40,42,73,83,86]. Some studies provided more detailed reliability evidence by reporting coefficients for individual subscales [46,83] or separately across population groups [48,75,90], allowing for a more nuanced interpretation of score consistency across specific dimensions and populations (see Table 3).

*3.3.1.2. Measurement stability: test-retest reliability.* Test-retest reliability consists of applying the same instrument on two occasions separated by a time interval, where the first point represents the test and the second point the retest. By measuring the correlation between test results and posttest, a reliability coefficient is established that represents the relationship between the variance of true scores on a test and the variance of observed scores [101]. These correlations are usually expressed in correlation coefficients such as

**Table 3**
Evaluation of internal consistency by instrument.

| Authors | Name tool | IC Test | Further details |
|---|---|---|---|
| Bubica & Boljat [42]. | Croatian CT assessment | α = .630 | Not reported |
| Basu et al. [40] | CT concepts and practices assessment instrument | CT practices α = .78 CT concepts, from α = .48 to α = .66 | Reliability was low for the three test factors of CT concepts but was high for CT practices. |
| Chen et al. [13] | Evaluation of CT of primary school students in everyday reasoning and robotic programming. | EAP/PV = .808 WLE = .979 | The instrument correctly ranked the students according to the construct. |
| Clarke-Midura et al. [44] | AT | KR-20 = .82 | Not reported |
| Coban & Korkmaz [46] | Computational Thinking Score (CTS) | α = .785 | For individual subscales the α values varied between .705 and .868 |
| El-Hamamsy et al. [48] | cCTt | α = .85 | High scores were preserved when discriminating by grades (3 and 4) (α = .84 in both cases). Reliability by question blocks was high for blocks 1–3 and 5 (.7< α < .9) and moderate for 4 and 6 (.5 < α < .7). |
| Kalyenci et al. [51] | CodingTest-A | KR-20 = .973 | Not reported |
| Khodabandelou & Alhoqani [52] | Nameless | α = .73 | Both in the pretest and the posttest. |
| Kong & Lai [53] | Nameless | α = .823 | Not reported |
| Kwon et al. [55] | Nameless | Pretest α = .70 Posttest α = .77 | Not reported |
| Li et al. [59] | CTA-CES | α = .76 | Not reported |
| Matere et al. [60] | Nameless | α = .77 | Not reported |
| Munawarah et al. [62] | Test for CT skills in the mathematics-based | α = .781 α = .709 | Multiple Choices items Essay items |
| Polat et al. [68] | CTt (Turkish version) | α = .80 | Not reported |
| Relkin et al. [73] | TechCheck | α = .68 | Not reported |
| Román-González et al. [75] | CTt | α = .793 | When discriminating by grades, high α scores were preserved (Fifth and sixth α = .721; Seventh and eighth α = .762; Ninth and tenth α = .824) |
| Sigayret et al. [80] | Nameless | α = .8 | Not reported |
| Sung [81] | Bebras Kards (Korean Adaptation) | α = .66 | Some subfactors had low α values (pattern recognition α = .52 and algorithms α = .49) |
| Sung [81] | TACTIC-KIBO (Korean Adaptation) | α = .88 | The α values of the individual subfactors ranged between α = .11 and .74, with the algorithm (α = .64) and representation (α = .74) subfactors being the highest. |
| Tran [83] | Nameless | Pretest α = .63 Posttest α = .61 | Not reported |
| Tsai et al. [85] | CTS | α = .91 overall scale | From .74 to .83 for the subscales |
| Tsai et al. [86] | CTT-ES | α = .69 EAP/PV = .69 WLE = .66 | Not reported |
| Wang et al. [87] | TMCYC | α = < .90 on all scales | Not reported |
| Zapata-Cáceres et al. [90] | BCTt | α = .824 | The higher the school grade, the lower the α value, so the authors concluded that the BCTt is mainly aimed at grades 1 to 4 of primary school, where it shows greater reliability. |
| Zhan et al. [91] | Nameless | α = .912 | Not reported |
| Zhang et al., [92] | | α = .774 | Not reported |
| Zhang & Wong [93] | CTt-LP | α = .873 | Not reported |
| Zhao [94] | MCTt | Pretest α = .74 Posttest α = .72 | Not reported |

Note: CI: Internal consistency; λ6: Guttman's Lambda 6; KR-20: Kuder-Richardson Formula 20 (KR20); α: Cronbach's alpha; WLE: Weighted Likelihood Estimate; EAP/PV: Expected A Posteriori estimate based on Plausible Values.

Spearman's correlation coefficient (ρ(ro)) and Pearson's correlation coefficient (r), both of which have values between −1 and 1 [102]. Benchmarks are usually mildly correlated from .0 to .20, fair from .21 to 40, moderate from .41 to .60, high from .61 to 80, and near perfect from .80 to 1 [101]. Four (6.2%) of the reviewed studies assessed test-retest reliability [8,78,87,90], all these studies found moderate to high stability in their assessment tools (See Table 4). We considered acceptable test-retest correlations > 50 (moderate to high).

*3.3.1.3. Alternate-form reliability.* Alternate-form reliability is like the test-retest (Table 5), but with the use of an alternate form of the same questionnaire. It is usually used when the test must be performed in a very short period, and you want to avoid an effect on the memory of the test-takers [96]. The coefficient used to indicate in which the instrument measures the same with both versions is called

**Table 4**
Evaluation of measurement stability by instrument.

| Authors | Name tool | Time Interval, Master | test-retest correlations | Further details |
|---|---|---|---|---|
| Rowe [78] | IACT | NR n = 1435 Between 3rd & 8th grade | Global scale r = .49 ($p < .0001$) Individual subscales between r = .18 and .40 ($p < .0001$) | The results indicate moderate stability in the global scale. |
| Wang et al. [87] | TMCYC | 2 weeks n = 15 Preschool | Between $r = .56$ ($p < .001$) and $r = .94$ ($p < .05$) | The results indicate moderate to high correlations. Only one game reported low correlation $r = .45$ ($p > .05$) |
| Zapata-Cáceres et al. [90] | BCTt | 5 weeks, n = 29 5th grade | ρ(ro) = .93 ($p < .01$) | The results indicate a very strong correlation, evidencing good stability. |
| Zhang & Wong [93] | CTt-LP test | 8 weeks n = 126 Between 1st & 3rd grade | ICC = .757 | The results show that the instrument allowed stable measurements. |

Note: ICC: The intraclass correlation coefficient; ρ (rho) = Spearman's correlation coefficient; r = Pearson's correlation coefficient.

**Table 5**
Evaluation of alternate-form reliability by instrument.

| Authors | Name tool | Further details |
|---|---|---|
| Kwon et al. [55] | NR | The study evaluated three parallel forms of the test calculating the Spearman-Brown coefficients, which ranged between .77 and .82, confirming the equivalence of the versions. |
| Chen et al. [13] | Evaluation of CT of primary school students in everyday reasoning and robotic programming. | This work evaluated two forms of the test, one based on programming language and the other based on blocks. A two-tailed *t*-test did not show a significant difference in performance in the two tests (t = 119; .556, $p$= .579), showing that the tests measured the same constructs. |
| Li et al. [59] | Computational Thinking Assessment for Chinese Elementary Students (CTA-CES) | This study evaluated six versions of the test with the same items but in a different order. The ANOVA showed that the mean values were significantly different between the versions, $F(5, 558) = 8.29$, $p < .001$. These results suggest that item position has an effect on test performance. The authors do not report an equivalence coefficient. |

**Table 6**
Evaluation of Reliability index of agreement between evaluators.

| Authors | Measurement | Inter-rater reliability | Additional Information |
|---|---|---|---|
| Angeli & Valanides [38] | Rubric for assessing children's CT | 97% agree | Reliability between two evaluators was established. The authors did not report the correlation coefficient or Cohen's/Fleiss' kappa. |
| Clarke-Midura [44] | AT | λ = .99 | Two evaluators independently evaluated 59 students with AT. |
| Clarke-Midura [45] | KRO indicators | λ = .828 | Agreement between evaluators, two researchers coded independently, showing strong agreement on the evaluation indicators. |
| Gane et al. [49] | Nameless | 3rd grade λ = .98 4th grade λ = .91 | They discriminated against the agreement index of those evaluated between third grade and fourth grade. |
| Metcalf et al. [61] | Functionality rubric | λ = .631 (69% agree) | Two evaluators applied the rubrics to 47 pairs of students, after resolving doubts from the evaluators about the processing, the rubrics were recoded, and a 100% agreement was reached. |
| | Concept Fluency Rubric | λ = .86 (90% agree) | Not reported |
| Relkin et al. [70] | TACTIC-KIBO | λ = 1 | Four external evaluators viewed the IPS video recordings of 14 students. The λ values between the primary rater and the guests ranged from moderate to strong. Variability in IPS scores was observed; however, the mean scores for each child across all raters were consistent with those of the primary evaluator. |
| Rowe et al. [77] | Zoombinis | λ = .70 | Two investigators independently labeled all rounds of the Level 1 game, most of which obtained scores above λ = .70 except for 3 labels. |
| de Ruiter et al. [37] | CSA | λ = .777 | Two raters independently rated 559 responses to the CSA from 23 children as satisfactory or unsatisfactory (videotaped sessions). |
| Wang et al. [87] | TMCYC | λ = .82 | The raters independently scored data from 15 children and interrater reliability was calculated. |

Note: r: Correlation coefficient.; λ: Fleiss ' Cohen Kappa.

the equivalence coefficient [103]. This method was used in three studies [13,55,59] (See Table 5).

*3.3.1.4. Split-half reliability.* Split-half reliability involves dividing an instrument into two parts to assess the consistency of its content rather than the temporal stability of the measurements. This is evaluated by calculating the correlation between both halves, with the most used statistical correlation being the Guttman's split-half correlation (λ6) [96]. Among the studies reviewed, only de Ruiter & Bers [37] applied this method to assess the Coding Stages Assessment (CSA), reporting a λ6 value of .94, indicating a high level of content consistency between the two halves of the instrument.

*3.3.1.5. Inter-rater reliability.* An important source of measurement error is the product of inter-observer variability (See Table 6), the magnitude of which can be estimated through concordance studies - it seeks to estimate the extent to which two observers agree in

**Table 7**
Evaluation of content validity.

| Author | Measurement | Number of experts | Additional Information |
|---|---|---|---|
| Basu et al. [40] | CT concepts assessment instrument | n = 5 | They assessed the alignment of each task with the objectives of the instrument, the suitability of the task for the grade level, and the clarity of each item, but did not provide statistical data in this regard. |
| Choi et al. [43] | Lee (1999) adaptation | n = 3 | Three primary school teachers reviewed the questionnaires and modified difficult phrases or expressions in the original assignments designed by Lee (1999). However, the authors do not report indices of agreement and do not provide methodological information on how the expert judgment was made. |
| Coban et al. [46] | Computational Thinking Score (CTS) | NA | They evaluated the suitability of each item through the feedback of a group of international experts; from this, the items were modified and passed to a second round of review by another external judge to define the final structure of the questionnaire. |
| El-Hamamsy et al. [48] | cCTt | n = 37 | They evaluated the level of difficulty. The informants reported a general average difficulty of .4 ± 1.0% and a progressive increase in difficulty throughout the test. |
| Kalyenci et al. [51] | CodingTest AB - Early Childhood Coding Skills Assessment Test | NA | They used the Lawshe technique to evaluate the judgment of a group of experts on the relevance of the CodingTest AB items; each item scored higher than .99. Therefore, corrections were made to some items and only one item that had a discrimination index of less than .19 was eliminated. |
| Kong & Lai [53] | CT concept test | n = 2 | Two expert judges (one aware of the research objectives and the other blind to them) review the instrument. The results of these reviews were positive and the authors pilot tested the instrument with primary school students. |
| Lee & Cho [56] | Computational Thinking Effectiveness test | n = 10 | They employed the Delphi method; the CVR values of all subareas were greater than .99, therefore, the validity of the contents of the instrument was confirmed. |
| Munawarah et al. [62] | Test for CT skills in the mathematics-based | n = 2 | In the adequacy of the item with the construct, they indicated that the results obtained from the CVR and CVI showed that 40 of the items support the validity of the test, therefore, it was reported that the prototype is valid. However, the authors do not report the values obtained. |
| Putri et al. [69] | Mathematical Computational Thinking Skill test (MCTS), | n = 7 | Through the V Aiken method, which expresses the content validity index through the "V coefficient", all the items of the MCTS test had a V coefficient of ≥ .76 ($p < .05$), the average of the entire test was .87; each of the scales also had important values: .87 decomposition, .88 pattern recognition, .86 abstraction, .83 algorithmic thinking, and .88 debugging. Therefore, it was concluded that MCTS is valid to measure the different factors established in the test. |
| Román-González et al. [4] | Computational Thinking test | n = 20 | Expert judges rated the level of difficulty, relevance, and acceptability of each item on the test. The judges recognized a clear trend of increasing difficulty throughout the test. The acceptability rate of the Items among the experts was 80 to 90%. Based on this review, the authors excluded some items, and the final version was made up of 28 items. |
| Ternik et al. [82] | Bebras Tasks | n = 5 | Five expert judges evaluated the classification of a set of Bebras tasks. The Fleiss' Kappa coefficient obtained for the five evaluators was k = .19, indicating a slight but weak level of agreement among them. |
| Wang et al. [87] | TMCYCs (Tool Measuring Coding Ability of Young Children) | n = 15 | They identified the item content validity index (I-CVI); 15 experts rated the importance and appropriateness of each item of TMCYC; each item obtained an I-CVI above .80, indicating good content validity. |
| Wu et al. [89] | NA | n = 3 | They used the Delphi method to analyze the judgment of 3 experts; however, the authors do not report the concordance index between them. |
| Zapata-Cáceres et al. [90] | BCTt | n = 45 | They estimated each item's relevance level and difficulty and there was an agreement of 68.3%. In this case, the judges perceived an increasing difficulty in the test. |
| Zhang & Wong [8] | CTt-LP | n = 20 | They rated each item according to how much the item aligned with the objective of the tool, the degree of difficulty and the clarity of the statements; in relation to how much they aligned with the constructed evaluated on a scale of 0 to 10, the items had values between 8.5 and 10 and the level of clarity scored between 7.5 and 9.5, while the level of difficulty was progressive. |

*Note:* I-CVI: item-level content validity index.

their measurement. This method is known as inter-rater reliability and consists of several raters rating the same subject and then estimating the inter-rater agreement index using tools such as Cohen's kappa coefficient [96].

The Cohen's Kappa coefficient reflects inter-observer agreement and can be calculated in tables of any dimension, as long as two observers are contrasted; in the case of three or more observers, the Fleiss' kappa coefficient is used [104]. The kappa coefficient can take values between −1 and +1. The closer the value is to +1, the greater the degree of inter-observer agreement [105].

Inter-rater reliability was evaluated in nine studies (14.1%), covering a total of 10 instruments (15.4%). Agreement between raters was primarily assessed using Fleiss' kappa (κ). Kappa values between .60 and .80 were considered indicative of acceptable agreement, whereas values above .80 indicated excellent agreement. As shown in Table 6, most instruments demonstrated acceptable to excellent levels of inter-rater agreement.

Only one study evaluated reliability using Item Response Theory (IRT), reporting the test information function rather than a conventional reliability coefficient. Zhang et al. [8] found that the maximum test information (14.65) was achieved at an ability level of .9, indicating that the CTtLP provides greater measurement precision for participants with slightly above-average ability while covering a broad range of ability levels. Because this evidence reflects conditional measurement precision rather than traditional reliability indices (e.g., internal consistency, test–retest, inter-rater, or alternate forms reliability), it was not classified within any of the reliability categories considered in this review.

### 3.3.2. *Validity*

Validity refers to the extent to which evidence and theory support the interpretations of test scores for their intended purposes. Consequently, the validation process involves gathering relevant evidence to establish a sound scientific foundation for such interpretations [36]. Various sources of validity evidence may contribute to evaluating a test, including: (1) content validity, which examines whether the items adequately represent the domain or skill the instrument is intended to measure; (2) criterion-related validity—particularly concurrent validity—which evaluates the degree to which test scores align with those obtained from another measure assessing the same construct; (3) construct validity, which examines whether test scores behave in accordance with theoretical expectations; and (4) face validity, which refers to the extent to which the instrument appears, from the perspective of the respondent, to measure what it claims to measure [96].

In this systematic review, at the study level (N = 64), the most frequently reported forms of validity evidence were criterion-related validity (26.6%, n = 17) and content validity (23.4%, n = 15), followed by construct validity (21.9%, n = 14) and face validity (9.4%, n = 6). Detailed results for each type of validity evidence are presented below.

*3.3.2.1. Content validity.* Content validity refers to the extent to which the elements of an instrument are relevant and representative of the intended construct for a specific assessment purpose [96]. The most widely used method for evaluating content validity is expert judgment, wherein specialists assess the importance, appropriateness, relevance, clarity, and difficulty of each item, as well as its contribution to the overall scale.

A common metric for quantifying content validity is the Content Validity Index (CVI), which has two forms: (1) the item-level content validity index (I-CVI), representing the proportion of experts who rate an item as highly relevant (scores of 3 or 4 on a predefined scale), and (2) the scale-level content validity index (S-CVI), which is derived either from the average I-CVI scores across all items (AVE) or through the universal agreement method (UA). A scale is considered to have excellent content validity if it achieves I-CVI values of $\geq$.78 and S-CVI values of $\geq$.80 [106].

Additionally, Lawshe's Content Validity Ratio (CVR) serves as a statistical method for determining the validity of individual items based on ratings from a panel of experts. This statistic is particularly useful for deciding whether to retain or discard specific items within an instrument [107].

Of the 15 studies reporting content validity (see Table 7), eight provided quantitative or otherwise interpretable results [4,8,51,56,69,82,87,90], whereas seven mentioned expert review without sufficiently reporting the procedure or level of agreement [40,43,46,48,53,62,89].

*3.3.2.2. Criterion validity.* Criterion-related validity refers to the extent to which scores obtained from an instrument are associated with an external criterion that measures the same or a closely related construct, thereby providing evidence of the instrument's ability to support predictions or decisions [108]. It is typically assessed by examining the relationship between scores on the instrument under evaluation (the predictor) and an independent criterion measure [109]. When both measures are administered within the same or a similar time frame, this evidence is commonly referred to as concurrent validity.

At the study level, 17 of the 64 included studies (26.6%) reported evidence of criterion-related validity [8,26,42,46,51,59,70,72,73,37,75–78,80,81,86]. These studies provided criterion-related validity evidence for 18 of the 65 instruments reviewed (27.7%) (see Table 8), as one study evaluated two separate instruments. In most cases, concurrent validity was examined by comparing scores from the instrument under evaluation with those obtained from an established measure assessing the same or a related construct. The studies generally reported correlation coefficients (r) to quantify the association between the two measures. Correlation coefficients greater than .60 were considered satisfactory, indicating a relatively strong association between the instrument and the criterion measure.

As can be seen in Table 8, some instruments showed strong criterion validity. Among them, CTt [75] and Early Childhood Coding Skills Assessment Test [51] showed high correlation with instruments that mediate problem solving, while TACTIC-KIBO [70] and TechCheck [73] showed a relationship with other performance tests that had already been previously validated. Finally, the

**Table 8**
Evaluation of criterion validity.

| Authors | Measurement | Comparison measurement | Correlation index | Additional Information |
|---|---|---|---|---|
| Bubica & Boljat [42] | Croatian CT assessment | Bebras | r = .495, p = .001 | The authors selected six Bebras tasks that showed moderate correlation with CT assessment. |
| Coban & Korkmaz [46] | Computational Thinking Score (CTS) Web App | Skill test | r = .776, p < .001 | The authors designed the Skill Test to be compared to the web application. It was made by 12 items being selected among the questions in Bebras UK and Bilge Kunduz. A highly positive relationship between the web application and the skill test. They also compared the performance in their web app with a Psychometric Scale -Computational thinking skill levels scale- [26]; the correlation between the two measures was low. |
| | | Psychometric Scale: Computational thinking skill levels scale | r = .230, p < .001 | |
| Kalyenci et al. [51] | Coding test | Problem solving scale (Oguz y Akyol, 2015) | r = .89 & r = .93, p < .01 | The scores obtained by 100 children in the two tests were compared. |
| Li et al. [59] | CTA-CES | Raven matrices | r = .47, p < .001 | They also correlated children's programming experience as a predictive element, finding that programming experience had a medium-sized effect on CTA-CES performance (t (186) = 4.24, p < .001, d = .58); however, the correlation between both factors was low. |
| | CTA-CES | Spatial ability | r = .43 p < .001 | |
| | CTA-CES | Reading comprehension | r = .54, p < .001 | |
| | CTA-CES | Programming experience | r = .18, p = .02 | |
| Moreno-León et al. [26] | Dr. Scratch | Human experts | r = .682, p < .0001 | The authors compared 317 evaluations conducted by experts with assessments automatically performed by the Dr. Scratch software. |
| Relkin et al. [70] | TACTIC-KIBO | Interactive Play Session (IPS) | r = .895, p < .001 | There was only a discrepancy in the evaluation of 4 subjects who qualified at level 4 in TACTIC-KIBO, but they were evaluated at level 3 in IPS, with a correlation between significant to very significant on the scale of both tests except for the design process scale. |
| Relkin et al. [73] | TechCheck | TACTIC-KIBO Relkin et al. [70] | r = .53, p < .001 | Moderate correlation was evidenced, but with high variability in the results of both instruments, especially among subjects with lower scores. |
| Relkin & Bers [72] | TechCheck-K | TechCheck | r = .76, p < .001 | The correct response pattern to the TechCheck-K (n = 89) was correlated with the response pattern observed in first graders (n = 288) and second graders (n = 480). |
| Román-González et al. [76] | CTt | Academic performance Experience course code.org | ρ(rho) = .47 | The authors found a correlation of ρ (rho) = .43 with the computer science course and ρ (rho) = .44 with the coding achievement in the code.org course. |
| | | Computational talents | d = .95 p = .026 | The predictive validity of the CTt in early differentiation between superior and average computational thinkers is demonstrated. |
| Román-González et al. [75] | CTt | Primary Mental Abilities Battery (PMA) | PMA total scale r = .540 p < .01 PMA-V r = .273 p < .01 PMA-N r = −.157 p < .01 PMA-S r = .439 p < .01 PMA-R r = .442 | CTt showed a moderate correlation with other measures such as the global PMA score, the reasoning factor (PMA-R) and special factor (PMA-S), indicating that CTt could be related to general mental ability (fluid intelligence); and, to a lesser extent, logical reasoning and spatial ability. It also showed a strong correlation with problem-solving ability measured with the RP30 test. |
| | CTt | Problem solving Test RP30 | r = .669; p < .01 | |
| Rowe et al. [77] | Automatic CT detector in Zoombinis | External CT assessments with digital, iterative and logical puzzles. | From r = .02 to .25 | The results of 714 and 918 students from third to eighth grades on both measures were compared. They found that the detectors, especially the problem-solving ones, were weakly related to other external test detectors. Except for one of the measures, the correlations were generally weak between the Zoombinis detectors and the external tests. |
| Rowe et al. [78] | IACT | Rubric Graded by Teachers (RGT) | Individual scales r = <.28 p < .0001 Aggregated CT r = .29, p < .0001 | The study was carried out with two samples (Sample 1 n = 1281 and Sample 1 n = 1742) taken at different times. The IACT results of the sample 1 were compared with an external measure consisting of the evaluation of teachers through a rubric. The results of sample 2 were compared with the results of 5 items of the Bebras task. Individual measures of IACT (Problem Decomposition, Pattern Recognition, Abstraction, Algorithm Design) were not strongly correlated with RGT or Bebras scores. However, aggregated CT (average across the 4 scales) showed a moderate correlation, providing some evidence that these measures assess the same construct. |
| | IACT | Bebras task | Individual scales r= < .15 p < .05 Aggregated CT .40, p < .0001 | |

(*continued on next page*)

**Table 8** (*continued*)

| Authors | Measurement | Comparison measurement | Correlation index | Additional Information |
|---|---|---|---|---|
| de Ruiter & Bers [37] | CSA | TechCheck [73] | r = .55, p < .01 | Moderate positive correlation indicating that the instruments would be partially evaluating the same construct. |
| Sigayret et al. [80] | Mastery of CT concepts and Algorithmic Test | Practice in programming | r = .58, p < .01 | The correlated scores of 449 students with different levels of programming practice concluded that the level of programming practice was positively correlated with performance on the test. |
| Sung [81] | Bebras Kards (Korean Adaptation) | TACTIC-KIBO (Korean Adaptation) | From r = .11 to r = .28 | Positive, though low, correlation between the total Bebras Kards scale and the TACTIC-KIBO Hardware, Software, Representation and Algorithms subscales |
| | TACTIC-KIBO (Korean Adaptation) | Bebras Kards (Korean Adaptation) | r = .18, p < .01 | TACTIC-KIBO total score is positively, but weakly, correlated to all the scales except for the Control Structure, Debugging, and Design Process subfactors. |
| Tsai et al. [86] | CTT-ES (result oriented) | CTS (process-oriented CT) | r = .274, p < .001 | The correlation coefficients were also weak, but significant (r = .154 to .291, p < .001). The results suggested that the factors of the two instruments are convergent.<br>The CTT-ES sub-scores were interrelated with the CTS sub-scores because each CTT-ES task was associated with more than one CT factor in CTS. |
| Zhang & Wong [8] | CTt-LP | Scratch performance evaluation. | r = .443, p < .000 | The researchers evaluated the correlation of the results of the CTt-LP with the scratch performance, evaluated by the teacher (n = 129). The results showed that the two measures were moderately correlated, indicating an acceptable criterion validity of CTtLP. |

*Note:* ρ (rho) = Spearman's correlation coefficient; r = Pearson's correlation coefficient.

Computational Thinking Score (CTS) software designed by Coban & Korkmaz [46] showed a high correlation with another instrument designed by the authors, based on Bebras UK and Bilge Kunduz.

Some assessment tools such as Croatian CT assessment [42], TechCheck-K [72], and CSA [37] showed a moderate correlation with other previously validated instruments, while CTA-CES [59] and CTt [75] showed a moderate correlation with instruments that mediate other cognitive variables. In some studies, the assessment instruments were moderately correlated with previous programming practice and experience, as was the case with the Mastery of CT concepts and Algorithmic Test [71] and CTt-LP [8]. We also observed some studies where the instruments showed a strong correlation with some measures, but a low correlation with measures that assessed perceived self-efficacy in programming [46] or with programming experience [59].

Four instruments showed a low correlation with other measures (r = .2 to r = .4), such as CTT-ES [86], IACT [78], the Korean adaptation of Bebras Kards and TACTIC-KIBO [81] and these last three instruments were correlated with measures based on Bebras.

In addition to concurrent validity, some studies have examined predictive validity. For instance, Tsai [86] conducted a hierarchical multiple regression analysis and found that the Abstraction, Decomposition, Evaluation, and Generalization scales were significant predictors of students' performance in CTT-ES. Similarly, Roman-Gonzalez et al. [75] assessed academic performance (computer science, math, and language) and Code.org performance at the beginning and end of the semester to test the predictive power of CTt. Their findings indicate that the instrument effectively identifies students with computational skills early in high school.

*3.3.2.3. Construct validity.* Construct validity indicates the extent to which an instrument's results behave in accordance with what is theoretically expected of it. This means that instrument scores should correlate with certain variables that are expected to be related and, at the same time, should not be related to others with which correlations are not expected [96]. One of the most common procedures to establish construct validity is Factor Analysis (exploratory and confirmatory), which examines the dimensionality between a set of items of measuring instruments by determining whether there is a greater relationship between responses within different subsets of items than between other subsets; the instrument is assumed to have construct validity when the result of this analysis coincides with the theoretical assumptions that support the instrument [110].

Exploratory factor analysis (EFA) makes it possible to elucidate the patterns of relationships between different items and constructs, as well as to identify potentially problematic elements that do not empirically belong to the expected construct. On the other hand, confirmatory factor analysis (CFA) seeks to verify the factor structure specified by the researcher; it is a popular method to support construct validity in measurement instruments [111].

Fourteen of the 64 included studies (21.9%) reported evidence of construct validity (see Table 9). Because one of these studies evaluated construct validity for two separate instruments, construct validity evidence was identified for a total of 15 of the 65 instruments reviewed (23.1%). Of the 14 studies, seven relied exclusively on exploratory factor analysis (EFA) [40,46,54,64,37,84,87], five relied exclusively on confirmatory factor analysis (CFA) [48,53,81,85,93], and two employed both exploratory and confirmatory factor analyses [66,86]. Regarding the factorial structure of the instruments, some supported a unidimensional structure [8,40,46,53,54,87], whereas others were better represented by multidimensional structures comprising two [64], three [81], four [54], five [84,86], or up to six factors [48,37,81].

Zhang and Wong [93] assessed the construct validity of the test using Item Response Theory (IRT) with the Three-Parameter

**Table 9**
Evaluation of construct validity.

| Authors | Measurement | Type of analysis | Findings |
|---|---|---|---|
| Basu et al. [40] | CT concepts and practices assessment instrument | EFA | In the CT practice test, the one-factor model demonstrated a good fit (RMSEA = .045, with RMSEA< .06). In the four-factor model, the constructs showed strong intercorrelations, ranging from .62 to .84. For the CT concept test, although the four-factor model exhibited a slightly better fit (RMSEA between .051 and .057), the one-factor model also provided an adequate fit (RMSEA of .061 or .062). These results suggest that it is appropriate to generate a composite score from the assessments at each level. |
| Cobán et al. [46] | CT Score (CTS) | EFA | The results suggested an unifactorial structure (ASC = .0125, Power Correlation .0003). A positive and significant correlation was observed between the seven skills that were intended to be assessed (Decomposition, Algorithmic Thinking, Abstraction, Pattern Recognition, Problem Solving, Critical Thinking, Creative Thinking), except between pattern recognition and abstraction ($p < .01$). However, the authors do not provide detailed statistics on these correlations. |
| El-Hamamsy et al. [48] | The competent Computational Thinking Test (cCTt) | CFA | Fit indices indicated an adequate model fit for the six factors (Sequences, Simple Loops, Complex Loops, Conditionals, While Statements, and Combinations): $\chi^2(260) = 551$, $p = .000$; TLI = .974; CFI = .978; RMSEA = .027; SRMR = .052. Factor loadings were positive and significant for all items (ranging from .540 to .867), with positive and significant correlations between factors. |
| Kong & Lai [53], | CT concepts test | CFA | The univariate model fit the data well ($\chi$ 2 (77) = 3000.312, $p < .001$; CFI = .965; TLI = .959; RMSEA = .054); All loads of the 14 items were positive with a value greater than .3. |
| Kong & Wang [54] | CT practices test | EFA | The four-dimensional model proved to be the most optimal (Testing and Debugging, Algorithmic Thinking, Reuse and Remixing, Abstraction and Modularization). After eliminating the 12 items with poor performance, the confirmatory MIRT analysis revealed that the bifactorial model with non-correlated factors had a better fit ($x^2 = 1092.31$, DF = 560, CFI = .95, TLI = .93, RMSEA = .04). All items charged highly and significantly in their corresponding latent factors. |
| Ortega-Ruipérez & Asensio [64] | Instrument to measure CT through complex problem solving. | EFA | The two-factor model (problem representation and problem solving) was the most appropriate (CFI = .979, TLI = .974, RMSEA = .017, p = .4325). The standardized correlation between the two factors was .969 (p = .000); which indicates a very good correlation. |
| Palts [66] | Bebras Challenger | EFA CFA | The four-factor model, encompassing Abstraction, Decomposition, Algorithmic Design, and Evaluation, demonstrated an adequate fit (TLI = .831, CFI = .865, RMSEA = .037); however, factor loadings were relatively low, ranging from .08 to .43. Given that the confirmatory factor analysis (CFA) did not yield the theoretically expected structure, an exploratory factor analysis (EFA) was conducted to identify the underlying dimensions. The EFA revealed a two-factor structure: (1) Algorithmic Design (F1), with factor loadings between .25 and .45, and (2) Pattern Recognition (F2), with loadings ranging from .20 to .29. A subsequent CFA confirmed the robustness of this two-dimensional model, demonstrating an excellent fit (TLI = 1.035, CFI = 1.000, RMSEA = .000, WRMR = .541). Factor loadings ranged from .429 to .745 for F1 and from .248 to .650 for F2, with a strong correlation of .677 between the two factors, supporting the validity of the revised model. |
| De Ruiter & Bers [37] | Coding Stages Assessment (CSA) | EFA | The Coding Skills Assessment (CSA) was developed based on Bers Coding Stages framework, which delineates six stages of coding skill development: Emergent, Coding and Decoding, Fluency, New Knowledge, Purpose and Determination. An exploratory factor analysis (EFA) indicated that a unidimensional model was suitable for three stages—Coding and Decoding, New Knowledge, and Purpose—whereas a two-factor model provided a better fit for the Emergent and Fluency scales. Correlation analysis among the items yielded an average correlation of $r = .24$ (range: from .07 to .39), suggesting that while the items exhibit reasonable homogeneity, they retain enough unique variance to avoid redundancy. |
| Sung [81] | Korean Adaptation of Bebras Kards | CFA | The three-factor model of the Bebras assessment (Korean adaptation)—comprising Patterns, Algorithms, and Logic—demonstrated a good fit ($\chi^2/df = 1.42$, $p < .001$; TLI = .83; CFI = .87; RMSEA = .03). The average variance extracted (AVE) index and composite reliability (CR) were .47 and .92, respectively, indicating acceptable convergent validity. However, some items with factor loadings below .30 were removed due to insufficient contribution to the model. |
| | Korean Adaptation of TACTIC-KIBO | CFA | The model demonstrated a good fit across its six factors—Hardware, Software, Representation, Control Structure, Pattern Recognition, Algorithm Design Process, and Debugging—($\chi^2/gl = 1.59$, $p = .00$, TLI = .80, CFI = .84, RMSEA = .03). Additionally, it exhibited strong convergent validity (AVE = .64, CR = .98). |
| Tsai [84] | Computational Thinking Scale (CTS) | EFA | The EFA results suggested grouping the 19 items into five dimensions—Decomposition, Abstraction, Algorithm, Evaluation, and Generalization—explaining 64% of the total variance. The square root of the Average Variance Extracted (AVE) exceeded .5 (range: .74 – .83) and was higher than the Pearson correlations between dimensions, confirming discriminant validity. |

(*continued on next page*)

**Table 9** (*continued*)

| Authors | Measurement | Type of analysis | Findings |
|---|---|---|---|
| | | | Additionally, all CTS dimensions showed significant positive correlations ($p < .001$), with coefficients ranging from .42 to .57, indicating moderate relationships between factors and supporting model acceptance. |
| Tsai [85] | Computational Thinking Scale (CTS) | CFA | A CFA confirmed all five factors, all item factor loads were greater than .70 (.73 to .86), and composite reliability values for each factor were greater than .70 (.84 to .89), AVE values for all factors were .57 to .67. The five variables were significantly correlated with each other (r = .42 to .58, $p < .001$). The square root of the AVE value of each factor was greater than .5 (from .75 to .82), confirming the discriminant validity. |
| Tsai [86] | CT Test for Elementary School Students (CTT-ES) | EFA | The exploratory factor analysis (EFA) results indicated that the 19 items clustered into five dimensions—Decomposition, Abstraction, Algorithm, Evaluation, and Generalization—accounting for 64% of the total explained variance. A confirmatory factor analysis (CFA) validated this five-factor structure, with all item factor loadings exceeding .70 (.73 to .86). Composite reliability values for each factor were above .70 (.84 to .89), while average variance extracted (AVE) values ranged from .57 to .67. The five dimensions were significantly correlated (r = .42 to .58, $p < .001$). Furthermore, the square root of the AVE for each factor was greater than .5 (.75 to .82), confirming discriminant validity. |
| Wang et al. [87] | TMCYC | EFA | The unifactorial model (Coding Capacity) showed a better fit, composed of nine items with an eigenvalue of 4.94. |
| Zhang & Wong [93] | CTt-LP | CFA | An unifactorial model showed adequate fit for the data set (RMSEA = .041, CFI = .911, TLI = .900). |

Note: RMSEA: Root Mean Square Error of Approximation; CFI: Comparative Fit Index; TLI Tucker-Lewis Index; χ2/gl Chi-Square Test/Degrees of Freedom; ASC Average Squared Correlation; $\chi^2$ Chi-Square; r Correlation Index; AVE: Average Variance Extracted; CR Composite Reliability; EFA Exploratory Factor Analysis; CFA Confirmatory Factor Analysis.

Logistic Model (3 PL), a psychometric approach designed to evaluate item quality and accurately estimate participants' abilities. Results indicated that the CTt-LP test showed the best fit among the models evaluated. All items met the monotonicity assumption and demonstrated acceptable fit indices, allowing for further calibration. Furthermore, no significant violations of local independence were observed. Regarding model parameters, items showed appropriate difficulty levels (M = .353), high discrimination (M = 2.188), and low guessing rates (below .35), ensuring precise measurement of performance. These findings provide strong evidence of the test's construct validity, confirming its suitability for the target population.

Based on the IRT analysis, the CTT-ES test demonstrated adequate model fit [86]. On the logit scale of the Rasch model, item difficulty for the CTT-ES ranged from 5.27 logits (highest) to −1.19 logits (lowest), with a mean competence level of .008 logits (SD = 1.13). Point-biserial correlations for correct responses yielded classical discrimination values between .09 and .55. Item-level fit to the Rasch model was adequate, with MNSQ values ranging from .92 to 1.12 (M = 1.00, SD = .06), within the acceptable range of .70–1.30.

*3.3.2.4. Face validity.* Face validity refers to the degree to which a measurement instrument appears to assess the intended construct, based on the perceptions of experts or users regarding the adequacy and relevance of its items. While not a rigorous psychometric indicator, it plays a crucial role in the instrument's acceptance and credibility [95].

**Table 10**
Evaluation of face validity.

| Authors | Measurement | Subjects | Findings |
|---|---|---|---|
| El-Hamamsy et al. [48] | cCTt | n = 37 experts | A total of 63% of experts agreed that the test was suitable for measuring CT skills in elementary school students, while 26% remained neutral, and 11% disagreed. |
| Khodabandelou & Alhoqani [52] | CTC | n = 15 students<br>n = 7 experts | As a result, some items were modified, leaving the final scale composed of 16 items. |
| Relkin [70] | TACTIC-KIBO | n = 15 students | To assess the apparent validity of TACTIC-KIBO and its suitability in age 5 to 7; however, the authors do not provide further details about the findings. |
| Relkin et al. [73] | TechCheck | n = 19 Subjects with CT experience (teachers and students) | To evaluate whether the items incorporated the variables for which they were designed and to assess their suitability for the target age groups. Fleiss' Kappa values indicated an inter-rater agreement of 81% (κ = .63 (95% CI), $p < .001$). |
| Rowe [78] | IACT | Students | It was tested through two rounds of iteration and interviews with students to see that writing elicited the CT practices of interest. |
| Sung [81] | Korean Adaptation of TACTIC-KIBO and Bebras kards. | n = 5 experts | The items accurately reflected the intended subdomains; however, they may have been too challenging for the target age groups. Additionally, some responses exhibited bias within the current sample, and two subfactors demonstrated low internal reliability (Hardware: α = .11; Debugging: α = .37). |

Face validity was reported in six studies (9.4%), covering seven instruments (10.8%) (see Table 10). The approaches used to assess face validity varied across studies. In some cases, the evaluation of item difficulty was categorized as evidence of content validity, whereas in others it was reported as face validity. Four studies assessed face validity by obtaining feedback from students [52,70,73,78], while three studies incorporated expert evaluations [48,52,81]. The face validity evidence reported across the reviewed instruments is summarized in Table 10.

### 3.3.3. Level of evidence for validity and reliability

In this systematic review, psychometric quality was assessed based on the evidence reported for the CT assessment instruments examined in each study. To maintain consistency with the primary unit of analysis, the overall frequency and distribution of satisfactory and unsatisfactory psychometric evidence were summarized at the study level (N = 64), while instrument-level counts (N = 65) are reported only when relevant, as one study evaluated two separate instruments. Overall, the evidence revealed considerable variability in both the extent and quality of psychometric reporting across studies. Twenty-two studies (34.4%) provided more than two types of satisfactory validity and/or reliability evidence, while an additional 24 studies (37.5%) reported at least one type of satisfactory psychometric evidence, most frequently internal consistency. In contrast, 14 studies (21.8%) did not report any validity or reliability evidence, and 11 studies (17.2%) presented at least one unsatisfactory psychometric result. Of these 11 studies, seven (10.9%) also reported at least one type of satisfactory evidence, whereas the remaining four (6.3%) reported only unsatisfactory evidence. These categories were not mutually exclusive, as individual studies could report both satisfactory and unsatisfactory results across different psychometric measures. These findings are summarized in Fig. 5, where green cells indicate satisfactory evidence, yellow cells represent unsatisfactory evidence, and gray cells denote the absence of reported evidence.

Furthermore, eight studies (12.5%) reported up to four different forms of validity and/or reliability evidence, reflecting a comparatively more comprehensive psychometric evaluation of the CT assessment instruments examined in these studies. Among these studies, the instruments receiving the broadest range of psychometric evaluation included the Computational Thinking Score (CTS) [46], the Competent Computational Thinking Test (cCTt) [48], the Computational Thinking Test (CTt) [4], the Computational Thinking Test for Lower Primary (CTt-LP) [92], the CTt for Beginners (BCTt) [90], the Tool Measuring Coding Ability of Young Children (TMCYC) [87], and the CT Concepts and CT Practices Assessment Instrument [40].

Three of these studies were conducted in Spain and examined instruments originally developed in Spanish [4,48,90], two were conducted in China [39,87], and one was conducted in Turkey [46]. Additionally, Sung [81] examined Korean adaptations of two instruments—Bebras Kards and TACTIC-KIBO—and reported four types of validity and reliability evidence. However, some of the psychometric results reported in this study, specifically those concerning internal consistency and criterion-related validity, were unsatisfactory.

Table 11 summarizes the target populations of the studies according to the psychometric evidence reported. It is important to note that some studies included assessment instruments intended for use across multiple educational levels. Although studies targeting preschool children were the least represented, most reported satisfactory validity and reliability evidence. In contrast, more than 20% of the studies targeting elementary and high school students either lacked sufficient psychometric evidence or reported unsatisfactory validity and/or reliability findings. These categories were not mutually exclusive, as studies could report both satisfactory and unsatisfactory psychometric evidence across different validity or reliability measures.

## 4. Discussion

This systematic review aimed to survey and critically examine the psychometric characteristics of performance-based assessment tools used to measure computational thinking (CT) in K–12 education. Prior reviews have documented growing academic interest in CT assessment as an emerging field of research [22,31], while also identifying several persistent challenges and limitations. These include a lack of consensus on the definition of CT and how it should be assessed [7,10], as well as a scarcity of tools designed for specific populations—such as students with learning disabilities, neurodevelopmental conditions, or diverse linguistic and cultural backgrounds [18,23,112]. This gap is particularly relevant for children with ADHD or ASD, learners with low literacy levels, emergent bilinguals, and students from rural or underserved educational contexts, who may require differentiated assessment formats or multimodal task designs to accurately capture their computational thinking skills. Furthermore, despite the widespread use of block-based programming environments such as Scratch, ScratchJr, Blockly, and Code.org in K–12 instruction, few instruments have been specifically designed to assess key CT concepts within these platforms [17]. Finally, the field still lacks consensus on evaluation criteria—namely, which CT components should be prioritized, how they should be operationalized, and what psychometric standards should guide instrument development [21,59].

**Table 11**
Level of evidence for validity and reliability.

| Level of evidence for validity and reliability | Preschool | Elementary school | High school |
|---|---|---|---|
| Total | N = 14 | N = 43 | N = 22 |
| No evidence | 2 (13%) | 8 (18%) | 5 (23%) |
| Unsatisfactory results | 1 (7%) | 4 (9%) | 1 (5%) |
| A satisfactory measure | 5 (33%) | 13 (30%) | 10 (45%) |
| Two or more satisfactory measures | 7 (47%) | 15 (34%) | 6 (27%) |

While previous reviews have documented general aspects of the validation processes employed, as well as the dimensions of validity and reliability examined [18,23,31], these studies do not offer a detailed analysis of the results obtained in each study or the degree to which the validity or reliability evidence of the reviewed instruments has been satisfactory. Consequently, the present study sought to update the state of the art in the assessment of CT in K-12 education, with a particular emphasis on the methodological soundness and psychometric characteristics of the available instruments. The validity and reliability of these tools were assessed in different educational levels, providing relevant information that might guide future research and improve the selection and employment of assessment instruments in educational and research contexts.

The instruments identified in this review varied substantially in format, application context, and the underlying CT model each was built to capture. Following the classifications proposed by Şenel and Şenel [113] and Weintrop et al. [15], CT assessment tools broadly fall into five categories: self-report scales measuring perceived ability; maximum-performance tests using multiple-choice or open-ended items; portfolio- or project-based assessments of accumulated artifacts; rubrics for scoring complex, open-ended performance; and technology-enhanced assessments that automatically analyze code, interaction logs, or other digital traces.

Within the performance-based subset examined here, task-based questionnaires dominated the field by a wide margin (n = 40, 62%; 51.5% multiple-choice), echoing the pattern reported in earlier reviews [22,31]. Rubrics for teacher observation were the second most common format (n = 21, 32.8%), followed at a distance by open-ended problem-solving tasks (n = 7, 11%) and, least frequently, software-based tools that score performance automatically within an application (n = 6, 9.4%). The scarcity of automated tools is notable given their potential for scalable, low-burden assessment; it is consistent, however, with earlier accounts of the accessibility and technical constraints that continue to limit their autonomous use for measuring CT [59].

This concentration in a small number of formats sits uneasily alongside the field's own recommendation that CT be assessed through complementary methods [22,75]: only ten of the 64 studies combined more than one assessment modality [13,41–43,56,58,59,78,37,83]. The gap is unsurprising — diversified assessment protocols are costly and logistically demanding to implement [22] — but it means that, for the large majority of instruments, CT competence is inferred from a single lens rather than triangulated across several.

That choice of lens appears to carry a psychometric cost. Task-based questionnaires and rubrics were, again, the two most-used formats [15,59], yet questionnaires accumulated markedly stronger validity evidence than rubrics did. The disparity likely reflects an inherent asymmetry between the two designs rather than any difference in researcher rigor: rubrics depend on observational, context-bound judgments that resist the kind of standardized validation questionnaires more readily accommodate [59].

Regarding the target population, instrument development in this field has been uneven across educational levels. Primary-level tools were the most numerous (67%), secondary-level tools considerably scarcer (34%), and preschool tools scarcer still (22%) — a gradient consistent with prior reviews [18,23,112] and one plausibly rooted in the difficulty of capturing CT reliably in very young children. Because CT curricula increasingly begin at the preschool level regardless, this imbalance leaves a meaningful portion of current instruction without age-appropriate instruments for determining how CT competence manifests at this developmental stage.

The type of instrument best suited to a given age group also appears to shift with the learner's autonomy. Direct performance measures grow more feasible as students require less scaffolding, whereas younger children continue to depend on teacher-mediated observation. This pattern surfaces clearly in the secondary-school data, where task-based questionnaires accounted for two-thirds of instruments (66%) and rubrics for barely a quarter (28%); the small number of software-based tools identified (N = 5) appeared exclusively at the primary and secondary levels, not in preschool assessment.

On the one hand, the lack of theoretical consensus on how computational thinking (CT) should be defined—and, consequently, which skills should be included within this construct—has resulted in multiple perspectives on how CT is assessed in children and adolescents in K–12 education. This concern has been consistently noted in previous literature [7,15,18,23], which argues that the still evolving and partially undefined nature of CT has led researchers to emphasize different dimensions when operationalizing the construct. Our systematic review reflects this same issue: several instruments reporting concurrent or criterion-related validity presented only low to moderate correlations with other CT measures, suggesting limited alignment in how CT components are conceptualized and measured across studies. These findings reinforce the call made by prior scholarship for the development of clearer frameworks that delineate what CT is, how its components should be defined, and how it differs from related cognitive or problem-solving constructs [6,10].

Despite the diversity of constructs that define CT, most of the reviewed assessment tools analyze algorithmic thinking as one of the main components to be assessed (66%), and even in some cases as the only CT construct [44,61,63,37,80]. This could be explained, as mentioned by Tang et al. [23], by the fact that most CT assessments focus on students' programming skills, and not, more broadly, on their CT skills.

Sixty-nine percent of the assessment tools relied on operational definitions [2,5] to establish the factors to be assessed in CT, such as abstraction, analysis, generalization, algorithmic thinking, and debugging. Similar findings were reported in the review by Zuñiga et al. [112], who explored empirical studies in which abstraction was the most frequent factor. Corrales-Álvarez et al. [18] and Cutumisu [22] consistently identified abstraction, algorithmic thinking, logical thinking, analysis, debugging, and modularization as the most frequently used factors.

On the other hand, a considerable proportion of instruments (39%) aimed to assess CT knowledge through educational or curriculum-based definitions [4]. Many of these tools explicitly draw on the theoretical framework proposed by Brennan and Resnick [10], which conceptualizes sequences, loops, events, and functions as core "computational thinking concepts," constructs that are foundational to algorithmic problem-solving. The prominence of this model in the instruments reviewed is consistent with findings from Ocampo et al. [31], whose systematic review of early-years CT education similarly identified Brennan and Resnick's three-dimensional framework as the most widely adopted. Together, these observations suggest that, despite broader conceptual fragmentation in the field, certain curricular definitions—particularly those associated with algorithmic structures—have gained

substantial traction in the operationalization and assessment of CT.

Finally, a subset of instruments (9.3%) assessed CT by prioritizing broader cognitive processes such as problem solving [46,57,77,91,114], critical thinking [46,114], logical reasoning [57], spatial thinking [44,30], and creative thinking [46,95]. Although these cognitive domains are theoretically related to computational thinking, Beecher [6] classifies them as superordinate or peripheral constructs, suggesting that—while relevant—they do not constitute core components of CT itself. This raises important questions about conceptual boundaries and underscores the need for clearer distinctions between CT and adjacent cognitive skills in assessment design.

Another factor contributing to the variability observed across instruments is the diversification of the contexts in which CT is taught—that is, the different pedagogical environments, modalities, and learning settings through which CT instruction is delivered. Importantly, CT learning is not limited to programming alone, which serves as one possible tool for fostering CT skills [5,23], but is only one among several instructional pathways. Waite [115] expands this discussion by identifying four situated pedagogical contexts for CT education: (1) physical computing, which involves robotics boards and tangible computing kits; (2) the creation of games and animations within block-based or embedded programming platforms; (3) unplugged instruction, which introduces CT concepts without digital devices; and (4) transversal or cross-curricular pedagogy, which uses CT ideas to support learning in other subject areas. These diverse instructional contexts shape the design of assessment instruments, contributing to the heterogeneity observed in the current review.

As noted above, 57% of the studies employed plugged activities, in which tasks or items involved the use of game-based environments, animations created with block-based programming languages, or robotics platforms. In contrast, 54% of the studies implemented assessment tasks that dispensed entirely with electronic devices, relying instead on unplugged activities. Only three studies [54,62,69] incorporated a cross-curricular approach to CT assessment, integrating computational thinking with mathematical problem-solving [62,69] or everyday reasoning tasks [54].

One of the main purposes of this review was to provide an overview of the level of validation of the available instruments. Unlike previous reviews, this study extracted and reported findings related to the validity and reliability of the instruments, presenting the evidence reported in the studies. Evaluating the validity and reliability of an assessment tool is critical to ensuring that the measures and information obtained are consistent and accurate, allowing for reliable evaluation [96].

Previous systematic reviews have revealed diverse data regarding the validity and reliability of CT assessment instruments. Tang et al. [23] found that 45% of the instruments used in empirical studies on CT reported reliability data, while only 18% included validity information. Corrales-Alvarez et al. [18] reported that 80% of the instruments used in studies with diverse populations presented evidence of reliability or validity. Ocampo et al. [31] found that, among the instruments designed to assess CT in young children, 54% reported evidence of validity and 88% reported reliability. Overall, these reviews have indicated that the internal consistency index is the most frequently reported indicator of reliability, while construct validity, followed by content and criterion validity, are the most documented types of validity.

The present review showed that 75% of the included studies reported at least one form of validity or reliability evidence, and 31% provided evidence for both. The discrepancy between these findings and those of previous systematic reviews can be explained by recent advances in the field as well as by the specific eligibility criteria employed in our search strategy. This review focused exclusively on instruments designed to assess CT performance skills in students from preschool through high school. Consequently, studies evaluating perceived self-efficacy or attitudinal measures were excluded, even though several of these instruments reported satisfactory validity and reliability evidence [85,116,117].

Likewise, instruments designed to assess other cognitive processes, or those linked to computing but not specifically to CT, were excluded [58,118–120]. To identify which populations have been targeted by instruments with the strongest evidence of reliability, studies were analyzed by age group. Although studies involving preschool populations were the least frequent (n = 14), most reported satisfactory psychometric evidence, with 85.7% (n = 12) reporting at least one satisfactory measure of validity and/or reliability. Among studies targeting primary school children (n = 43), 65.1% (n = 28) reported at least one satisfactory measure of validity and/or reliability. Among studies targeting primary school children (n = 43), 74% reported evidence of reliability or validity, of which 65% (n = 28) yielded satisfactory results.

Finally, at the secondary education level (n = 22), 72.7% (n = 16) of the studies reported at least one satisfactory measure of validity and/or reliability. Overall, these findings suggest that studies targeting preschool and school-aged populations generally reported satisfactory psychometric evidence for the performance-based CT assessments examined. However, the number of studies reporting satisfactory psychometric evidence was higher at the primary school level (n = 28) than at the secondary (n = 16) and preschool (n = 12) levels. This distribution warrants attention from the academic community, as the assessment of CT skills may become increasingly relevant—and potentially more challenging to operationalize—as students progress through secondary education. The comparatively smaller number of studies reporting satisfactory psychometric evidence at the secondary level may reflect the greater complexity of CT skills at this developmental stage, highlighting the need for further research on the assessment approaches and underlying CT components most appropriate for this population.

### *4.1. Reliability evidence reported in the studies*

Internal consistency was the most frequently reported reliability index, documented in 27 of the 64 studies (42.2%), of which 21 met conventional acceptability thresholds and five fell below .70 [73,81,83,86]. That this proportion held equally for questionnaires and rubrics suggests internal consistency functions as a default check computable from a single administration, rather than an index specifically suited to either format.

The remaining reliability indices tracked far more closely with instrument type. Inter-rater reliability was reported in only nine of

the 21 studies using rubric-based assessments [38,44,45,49,61,70,77,37,87]. and concentrated heavily in preschool assessments (n = 5), where teacher-mediated scoring is standard. For rubrics used with older students, inter-rater evidence was largely absent, leaving scorer subjectivity — arguably the greatest threat to validity in rubric-based CT assessment — underexamined precisely where it matters most.

Test-retest, parallel-forms, and split-half evidence were each confined to a handful of task-based questionnaires: four studies reported test-retest correlations, two using open-ended items [78,87] and two using multiple-choice items [8,90]; three used parallel forms [13,55,58]; and one used split-half estimation [37]. All reported values fell within acceptable ranges, but none of these three indices was applied to a single rubric, portfolio, or software-based instrument, so their temporal and structural stability remains, in effect, untested.

### 4.2. Validity evidence reported in the studies

Content validity was reported in 15 studies (23.4%), corresponding to 15 instruments (23.1%). Even among these, the underlying procedure was often left vague: nine studies specified a concrete method, such as the Lawshe technique [51], the Delphi method [56, 89], or Aiken's V [69], with an explicit coefficient reported in only two cases [56,87], while the remaining six offered little beyond a general mention of expert review. Content validity, in other words, appears to be treated more as a procedural formality than as a result worth reporting in detail.

Criterion validity was only marginally more common (26%) [8,26,42,46,51,59,70,72,73,37,75–78,80,81,86], most often operationalized as concurrence with an established CT measure [8,72,73,78,37], with indicators of programming experience [8,59,80], or with related constructs such as fluid intelligence [75] or reading comprehension [59]. This modest rate is plausibly a symptom of the field's youth: with few CT instruments yet established as reference standards, researchers have limited options against which to benchmark a new tool. Criterion-related validity was reported for 18 instruments across 17 studies (26.6% of the included studies). The magnitude of the associations varied considerably, ranging from weak to strong, and several instruments showed only limited convergence with the selected criteria [77,78,81].

Construct validity — arguably the most theoretically consequential form of evidence, given that most instruments were designed to capture multiple CT dimensions — was also the least pursued (22%). The gap was sharpest by format: thirteen of the studies examining factorial structure used task-based questionnaires, one of which also incorporated open-ended tasks; one additional study validated an open-ended performance task exclusively.

Among the 14 studies reporting construct validity, seven relied exclusively on EFA, five exclusively on CFA, and two employed both approaches, a consistent pattern emerged: instruments conceived as multidimensional tended, in practice, to fit better as unidimensional [8,40,46,53,54,37] — suggesting that the skills comprising CT may be more tightly interrelated than current theoretical models assume, although this could equally reflect limited statistical power to detect separate subfactors. Only four instruments reproduced their originally hypothesized structure: three built on Wing's and Selby's frameworks [54,81,84], and one on Brennan and Resnick's model of sequences, loops, conditionals, events, and functions [10]. Item Response Theory remains almost entirely unused in this field, applied in only two of the reviewed studies — Zhang and Wong's CTt-LP [93] and Tsai's CTT-ES [86] — both of which reported adequate model fit but leave open how CT items behave across the wider ability spectrum.

Face validity, the weakest form of validity evidence in classical test theory, was correspondingly the least examined of all four types (9.4%) [48,52,70,73,78,81]. Every study that did assess it, however, reported satisfactory results — a uniformity that may reflect genuinely well-constructed items, though it could equally point to the limited discriminating power of face validity as a psychometric check.

## 5. Conclusions

Computational thinking assessment in K–12 education remains a young and fast-moving area of pedagogical and psychological research, one that has grown markedly over the past decade as instrument design and validation studies have proliferated [24]. Across this growing body of work, this review found recurring strengths — most notably a shared emphasis on algorithmic thinking and a common core of CT components — alongside six challenges that, taken together, constrain confidence in the field's assessment tools: (1) construct validity remains rarely examined, with evidence reported in only 14 studies (21.9%), covering 15 instruments (23.1%); (2) content validity is often asserted rather than demonstrated, with expert-judgment procedures left undocumented in many studies; (3) face validity was reported in only six studies (9.4%), covering seven instruments (10.8%); (4) criterion and concurrent validity, where reported, tend toward low-to-moderate correlations and uneven documentation; (5) advanced psychometric approaches such as Item Response Theory remain largely untapped; and (6) reliability evidence for rubric- and software-based instruments is sparse, with inter-rater agreement and internal consistency rarely reported for these formats.

These gaps point to a field still lacking a shared framework for both identifying appropriate CT measures and building more rigorous evidence around them [41,47,57,30,62–67,74,79,88,95]. Without such a framework, the evidentiary weight any single instrument can bear remains limited, with consequences for both research conclusions and the educational decisions built on them. This review helps narrow that gap by identifying the instruments with the strongest psychometric support currently available, while making explicit where the evidence base still falls short.

Much of this uncertainty traces back to a more basic problem: the field has yet to agree on what CT comprises, and the diversity of components targeted across the reviewed instruments is as much a symptom of that disagreement as a design choice. A related and largely separate gap concerns geographic and linguistic reach. North America accounted for 35% of the studies and instruments

reviewed, echoing the concentration already noted in prior reviews [18] and leaving Central and South America comparatively underrepresented. English-language instruments predominated (44%), with Spanish a distant second (13%) and smaller numbers in Korean, Turkish, Chinese, Japanese, French, and Arabic — a distribution that argues for deliberate investment in assessment tools built for other linguistic and cultural contexts, rather than translation as an afterthought. A few instruments, including the Bebras cards [121], the CT-cube/Cross Array Task [67], and the CTt for Beginners [90], have already been validated across broader populations and may offer a practical starting point for that work. Longitudinal validity — whether an instrument can track CT development over time, not just capture a single snapshot — remains almost entirely unaddressed and is a further gap worth prioritizing.

The uneven age coverage identified throughout this review bears repeating as a standalone concern: instrument development has concentrated heavily on primary school students, leaving preschool and secondary populations comparatively underserved. Closing this gap will require assessment tools built specifically for each developmental stage, rather than instruments adapted after the fact, if CT is to be measured consistently across the full K–12 continuum.

Two limitations of this review itself are worth noting. First, given how quickly this literature is expanding, the evidence synthesized here reflects a search conducted in 2023; periodic updates will be necessary to keep these findings current, and the review protocol has been published previously to support exactly that kind of replication. Second, this review did not distinguish original instruments from adaptations of existing tools. Future work would benefit from evaluating these two categories separately, since the evidentiary expectations for a novel instrument arguably differ from those for an adaptation building on an already-validated foundation.

Taken together, these findings make the case that CT assessment warrants more sustained methodological attention than it has so far received. As computational thinking continues to consolidate its place in education systems worldwide, the tools used to measure it need to keep pace — through rigorous, psychometrically sound development rather than through proliferation alone. By identifying which instruments currently offer the strongest evidence, and where that evidence remains thin, this review aims to give the field a clearer basis for building the next generation of CT assessment tools.

## CRediT authorship contribution statement

**Carolina Robledo-Castro:** Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization, Writing – review & editing. **Gabriele Pozzan:** Investigation, Formal analysis, Data curation, Conceptualization, Validation, Writing – review & editing. **Chiara Montuori:** Writing – review & editing, Investigation, Formal analysis, Data curation, Validation, Writing – original draft. **Tullio Vardanega:** Writing – review & editing, Data curation, Formal analysis, Investigation, Validation. **Barbara Arfè:** Writing – review & editing, Validation, Supervision, Conceptualization, Methodology.

## Data availability statement

Data will be made available upon reasonable request. For requests regarding the data supporting this study, please contact the corresponding author.

## Declaration on the use of artificial intelligence

During the preparation of this work, the authors used ChatGPT Plus (OpenAI) solely to improve the clarity, grammar, and readability of the manuscript. After using this tool, the authors carefully reviewed, edited, and verified all the content and take full responsibility for the accuracy and integrity of the manuscript. No generative AI or AI-assisted tools were used to generate, analyze, or interpret the data, prepare the study results or conclusions, or create or modify any figures or images included in this manuscript.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Carolina Robledo-Castro reports financial support was provided by Colombia Ministry of Science Technology and Innovation. Carolina Robledo-Castro reports financial support was provided by University of Tolima. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This study was conducted as part of the doctoral research internship undertaken by the first author at the University of Padova within her doctoral training in Cognitive Sciences at the Universidad Autónoma de Manizales. The authors sincerely thank Professor Luis Fernando Castillo Ossa, Ph.D., for his valuable guidance and support throughout the research process as doctoral thesis advisor. This work was supported by the Bicentennial Doctoral Excellence Scholarship Program of the Colombian Ministry of Science, Technology and Innovation (Minciencias) and by the University of Tolima. The research was conducted within the framework of the

Currículo, Universidad y Sociedad research group at the University of Tolima and in collaboration with the CoThi research group at the University of Padova.

## References


[1] J.M. Wing, Computational thinking, Commun. ACM 49 (3) (2006) 33–35. https://www.cs.cmu.edu/afs/cs/Web/People/15110-s13/Wing06-ct.pdf.
[2] J.M. Wing, Computational thinking and thinking about computing, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 366 (1881) (2008) 3717–3725, https://doi.org/10.1098/rsta.2008.0118.
[3] A.V. Aho, Computation and computational thinking, Comput. J. 55 (7) (2012) 832–835, https://doi.org/10.1093/comjnl/bxs074.
[4] M.R. Román-González, Computational thinking test: Design guidelines and content validation, in: Proceedings of EDULEARN15 Conference, Barcelona, Spain, 2015, pp. 2436–2444, https://doi.org/10.13140/RG.2.1.4203.4329.
[5] V.J. Shute, C. Sun, J. Asbell-Clarke, Demystifying computational thinking, Educ. Res. Rev. 22 (2017) 142–158, https://doi.org/10.1016/j.edurev.2017.09.003.
[6] K. Beecher. Computational thinking: a beginner's guide to problem-solving and programming, BCS, The Chartered Institute for IT, Swindon, 2017.
[7] C. Selby, J. Woollard, Computational thinking: the developing definitions, in: Proc. 45th ACM Tech. Symp. Comput. Sci. Educ., ACM, Canterbury, 2013, https://doi.org/10.1145/3159450.3162325.
[8] S. Zhang, G.K.W. Wong, Understanding individual differences in computational thinking development of primary school students: a three-wave longitudinal study, J. Comput. Assist. Learn. 40 (6) (2024) 2604–2615.
[9] Computer Science Teachers Association CSTA, International Society for Technology in Education ISTE, Operational Definition of Computational Thinking for K–12 Education, CSTA & ISTE, 2011.
[10] K. Brennan, M. Resnick, New frameworks for studying and assessing the development of computational thinking, in: Proc. Annual Meeting of the, American Educational Research Association, Vancouver, Canada, 2012.
[11] Q. Burke, W.I. Byrne, Y.B. Kafai, Computational participation: understanding coding as an extension of literacy instruction, J. Adolesc. Adult Literacy 59 (4) (2016) 371–375, https://doi.org/10.1002/jaal.496.
[12] A. Labusch, B. Eickelmann, M. Vennemann, Computational thinking processes and their congruence with problem-solving and information processing, in: S. C. Kong, H. Abelson (Eds.), Computational Thinking Education, Springer, Singapore, 2019, pp. 65–84, https://doi.org/10.1007/978-981-13-6528-7_5.
[13] G. Chen, J. Shen, L. Barth-Cohen, S. Jiang, X. Huang, M. Eltoukhy, Assessing elementary students' computational thinking in everyday reasoning and robotics programming, Comput. Educ. 109 (2017) 162–175, https://doi.org/10.1016/j.compedu.2017.03.001.
[14] S.C. Kong, H. Abelson, M. Lai, Introduction to computational thinking education, in: S.C. Kong, H. Abelson (Eds.), Computational Thinking Education, Springer, Singapore, 2019, pp. 1–10, https://doi.org/10.1007/978-981-13-6528-7_1.
[15] D. Weintrop, D. Rutstein, M. Bienkowski, S. McGee, Assessment of computational thinking, in: Computational Thinking in Education, Routledge, London, 2021.
[16] S. Bocconi, A. Chioccariello, G. Dettori, A. Ferrari, K. Engelhardt, P. Kampylis, Y. Punie, Exploring the field of computational thinking as a 21st century skill, Proc. EDULEARN16 (2016), https://doi.org/10.21125/edulearn.2016.2136.
[17] S. Grover, S. Cooper, R. Pea, Assessing computational learning in K-12, in: Proc. 2014 Conf. Innovation and Technology in Computer Science Education, ACM, 2014, pp. 57–62.
[18] M. Corrales-Alvarez, L.M. Ocampo, S.A. Cardona, Instruments for evaluating computational thinking: a systematic review, Tecnológicas 27 (59) (2023) 1–26, https://doi.org/10.22430/22565337.2950.
[19] J. Guggemos, S. Seufert, M. Román-González, Computational thinking assessment—towards more vivid interpretations, Technol. Knowl. Learn. 28 (2022) 539–568, https://doi.org/10.1007/s10758-021-09587-2.
[20] M. Babazadeh, L. Negrini, How is computational thinking assessed in European K-12 education? A systematic review, Int. J. Comput. Sci. Eng. Syst. 5 (2022) 3–19, https://doi.org/10.21585/ijcses.v5i4.138.
[21] N. da Cruz Alves, C. Gresse von Wangenheim, J. Hauck, Approaches to assess computational thinking competences based on code analysis in K-12 education: a systematic mapping study, Inf. Educ. 18 (2019) 17–39, https://doi.org/10.15388/infedu.2019.02.
[22] M. Cutumisu, C. Adams, C. Lu, A scoping review of empirical research on recent computational thinking assessments, J. Sci. Educ. Technol. 28 (6) (2019) 651–676, https://doi.org/10.1007/s10956-019-09799-3.
[23] X. Tang, Y. Yin, Q. Lin, R. Hadad, X. Zhai, Assessing computational thinking: a systematic review of empirical studies, Comput. Educ. 148 (2020) 103798, https://doi.org/10.1016/j.compedu.2019.103798.
[24] J. Lockwood, A. Mooney, Computational thinking in education: where does it fit? A systematic literature review, Int. J. Comput. Sci. Eng. Syst. 2 (2017), https://doi.org/10.48550/arXiv.1703.07659.
[25] V. Dagienė, G. Futschek, Bebras international contest on informatics and computer literacy: criteria for good tasks, in: R.T. Mittermeir, M.M. Sysło (Eds.), Informatics Education – Supporting Computational Thinking, Lecture Notes in Computer Science, 5090, Springer, Berlin, Heidelberg, 2008, pp. 19–30, https://doi.org/10.1007/978-3-540-69924-8_2.
[26] J. Moreno-León, M. Román-González, C. Harteveld, G. Robles, On the automatic assessment of computational thinking skills: a comparison with human experts, in: Proc. CHI Conf. Hum. Factors Comput. Syst, Extended Abstracts. ACM, 2017, pp. 2788–2795, https://doi.org/10.1145/3027063.3053216.
[27] Ö. Korkmaz, R. Çakir, M.Y. Özden, A validity and reliability study of the computational thinking scales (CTS), Comput. Hum. Behav. 72 (2017) 558–569, https://doi.org/10.1016/j.chb.2017.01.005.
[28] C. Pei, D. Weintrop, U. Wilensky, Cultivating computational thinking practices and mathematical habits of mind in lattice land, Math. Think. Learn. 20 (1) (2018) 75–89, https://doi.org/10.1080/10986065.2018.1403543.
[29] C. Jenkins, Poem generator: a comparative quantitative evaluation of a microworlds-based learning approach for teaching English, Int. J. Educ. Dev. using Inf. Commun. Technol. (IJEDICT) 11 (2) (2015) 153–167.
[30] J. Moreno-León, G. Robles, M. Román-González, Dr. Scratch: automatic analysis of scratch projects to assess and foster computational thinking, Rev. Educ. Distancia. 46 (2015) 1–23, https://doi.org/10.6018/red/46/10.
[31] L.M. Ocampo, M. Corrales-Álvarez, S.A. Cardona, Systematic review of instruments to assess computational thinking in early years of schooling, Educ. Sci. 14 (10) (2024) 1124, https://doi.org/10.3390/educsci14101124.
[32] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, D. Moher, The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, Syst. Rev. 10 (1) (2021) 89, https://doi.org/10.1186/s13643-021-01626-4.
[33] A.M. Methley, S. Campbell, C. Chew-Graham, R. McNally, S.P.I.C.O. Cheraghi-Sohi, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews, BMC Health Serv. Res. 14 (2014) 579, https://doi.org/10.1186/s12913-014-0579-0.
[34] C. Robledo Castro, T. Vardanega, G. Pozzan, C. Montuori, B. Arfè, Characteristics and psychometric properties of computational thinking assessments in children and adolescents: a systematic review, INPLASY Protoc (2023) 202340069, https://doi.org/10.37766/inplasy2023.4.0069.
[35] I. Boutron, M.J. Page, J.P.T. Higgins, D.G. Altman, A. Lundh, A. Hróbjartsson, D. Moher, Considering bias and methodological limitations when interpreting evidence from systematic reviews and meta-analyses, J. Clin. Epidemiol. 158 (2023) 1–9, https://doi.org/10.1016/j.jclinepi.2023.01.001.
[36] American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Standards for Educational and Psychological Testing, American Educational Research Association, Washington, DC, 2014.
[37] L.E. de Ruiter, M.U. Bers, The coding stages assessment: development and validation of an instrument for assessing young children's proficiency in the ScratchJr programming language, Comput. Sci. Educ. 32 (4) (2022) 388–417, https://doi.org/10.1080/08993408.2021.1956216.

[38] C. Angeli, N. Valanides, Developing young children's computational thinking with educational robotics: an interaction effect between gender and scaffolding strategy, Comput. Hum. Behav. 105 (2020) 106234, https://doi.org/10.1016/j.chb.2019.03.018.

[39] S. Basu, J.S. Kinnebrew, G. Biswas, Assessing student performance in a computational-thinking-based science learning environment, in: S. Trausan-Matu, K. E. Boyer, M. Crosby, K. Panourgia (Eds.), Intelligent Tutoring Systems. Lecture Notes in Computer Science, 8474, Springer, Cham, 2014, pp. 476–481, https://doi.org/10.1007/978-3-319-07221-0_59.

[40] S. Basu, D. Rutstein, Y. Xu, H. Wang, L. Shear, A principled approach to designing computational thinking concepts and practices assessments for upper elementary grades, Comput. Sci. Educ. 31 (2021) 1–30, https://doi.org/10.1080/08993408.2020.1866939.

[41] M.U. Bers, The TangibleK robotics program: applied computational thinking for young children. Early child, Res. Pract. 12 (2) (2010) 1–20. http://ecrp.uiuc.edu/v12n2/bers.html.

[42] N. Bubica, I. Boljat, Assessment of computational thinking: a Croatian evidence-centered design model, Inf. Educ. 21 (3) (2022) 425–463, https://doi.org/10.15388/infedu.2022.17.

[43] J. Choi, Y. Lee, E. Lee, Puzzle-based algorithm learning for cultivating computational thinking. Wirel. Pers, Commun. Now. 93 (1) (2017) 131–145, https://doi.org/10.1007/s11277-016-3679-9.

[44] J. Clarke-Midura, D. Silvis, J.F. Shumway, V.R. Lee, J.R. Kozlowski, Developing a kindergarten computational thinking assessment using evidence-centered design: the case of algorithmic thinking, Comput. Sci. Educ. 31 (2) (2021) 117–140, https://doi.org/10.1080/08993408.2021.1877988.

[45] J. Clarke-Midura, V.R. Lee, J.F. Shumway, D. Silvis, J.S. Kozlowski, R. Peterson, Designing formative assessments of early childhood computational thinking. Early child, Res. Q. 65 (2023) 68–80, https://doi.org/10.1016/j.ecresq.2023.05.013.

[46] E. Coban, Ö. Korkmaz, An alternative approach for measuring computational thinking: performance-Based platform, Think. Ski. Creat. 42 (2021) 100929, https://doi.org/10.1016/j.tsc.2021.100929.

[47] V. Dagienė, G. Stupurienė, Bebras: a sustainable community building model for concept-based learning of informatics and computational thinking, Inf. Educ. 15 (1) (2016) 25–44, https://doi.org/10.15388/infedu.2016.02.

[48] L. El-Hamamsy, M. Zapata-Cáceres, E.M. Barroso, F. Mondada, J.D. Zufferey, B. Bruno, The competent Computational thinking test: development and validation of an unplugged computational thinking test for upper primary school, J. Educ. Comput. Res. 60 (7) (2022) 1719–1756, https://doi.org/10.1177/07356331221088906.

[49] B.D. Gane, M.I. Noor, W.Y. Feiya, J.W. Pellegrino, Design and validation of learning trajectory-based assessments for computational thinking in upper elementary grades, Comput. Sci. Educ. 31 (2) (2021) 141–168, https://doi.org/10.1080/08993408.2021.1874221.

[50] D. Hooshyar, L. Malva, Y. Yang, M. Pedaste, M. Wang, H. Lim, An adaptive educational computer game: effects on students' knowledge and learning attitude in computational thinking, Comput. Hum. Behav. 114 (2021) 106575, https://doi.org/10.1016/j.chb.2020.106575.

[51] D. Kalyenci, Ş. Metin, M. Başaran, Test for assessing coding skills in early childhood, Educ. Inf. Technol. 27 (2022) 4685–4708, https://doi.org/10.1007/s10639-021-10803-w.

[52] R. Khodabandelou, K. Alhoqani, The effects of WeDo 2.0 robot workshop on Omani grade 5 students' acquisition of computational thinking concepts and acceptance of robot technology, Educ. Next 3 (13) (2022), https://doi.org/10.1080/03004279.2022.2041685.

[53] S.C. Kong, M. Lai, Validating a computational thinking concepts test for primary education using item response theory: an analysis of students' responses, Comput. Educ. 187 (2022) 104562, https://doi.org/10.1016/j.compedu.2022.104562.

[54] S.C. Kong, Y. Wang, Item response analysis of computational thinking practices: test characteristics and students' learning abilities in visual programming contexts, Comput. Hum. Behav. 122 (2021) 106836, https://doi.org/10.1016/j.chb.2021.106836.

[55] K. Kwon, A.T. Ottenbreit-Leftwich, T.A. Brush, et al., Integration of problem-based learning in elementary computer science education: effects on computational thinking and attitudes, Educ. Technol. Res. Dev. 69 (2021) 2761–2787, https://doi.org/10.1007/s11423-021-10034-3.

[56] K. Lee, J. Cho, Computational thinking evaluation tool development for early childhood software education, Int. J. Inform. Vis. 5 (3) (2021) 313–319, https://doi.org/10.30630/joiv.5.3.672.

[57] J. Leonard, A. Buss, R. Gamboa, et al., Using robotics and game design to enhance children's self-efficacy, STEM attitudes, and computational thinking skills, J. Sci. Educ. Technol. 25 (2016) 860–876, https://doi.org/10.1007/s10956-016-9628-2.

[58] T. Li, A. Traynor, The use of cognitive diagnostic modeling in the assessment of computational thinking, AERA Open 8 (2022), https://doi.org/10.1177/23328584221081256.

[59] Y. Li, S. Xu, J. Liu, Development and validation of computational thinking assessment of Chinese elementary school students, J. Pac. Rim Psychol. 15 (2021), https://doi.org/10.1177/18344909211010240.

[60] I. Matere, C. Weng, M. Astatke, C.H. Hsia, C.G. Fan, Effect of design-based learning on elementary students' computational thinking skills in visual programming maker course, Interact. Learn. Environ. (2021) 1–14, https://doi.org/10.1080/10494820.2021.1938612.

[61] S.J. Metcalf, J.M. Reilly, S. Jeon, A. Wang, A. Pyers, K. Brennan, C. Dede, Assessing computational thinking through the lenses of functionality and computational fluency, Comput. Sci. Educ. 31 (2) (2021) 199–223, https://doi.org/10.1080/08993408.2020.1866932.

[62] Munawarah, S.Z. Thalhah, A.D. Angriani, F. Nur, A. Kusumayanti, Development of instrument test computational thinking skills IJHS/JHS based RME approach, Math. Teach. Res. J. 13 (4) (2021) 202–220.

[63] Y. Oomori, H. Tsukamoto, H. Nagumo, Y. Takemura, K. Iida, A. Monden, K. Matsumoto, Algorithmic expressions for assessing algorithmic thinking ability of elementary school children, in: Proc. IEEE Front. Educ. Conf. (FIE). IEEE, 2019, pp. 1–8, https://doi.org/10.1109/FIE43999.2019.9028486.

[64] B. Ortega-Ruipérez, M. Asensio, Evaluating computational thinking through problem solving: validation of an assessment instrument, Ibero-Am. J. Educ. Eval 14 (1) (2021) 153–171, https://doi.org/10.15366/riee2021.14.1.009.

[65] G. Ota, Y. Morimoto, H. Kato, Ninja code village for scratch: function samples, function analyser and automatic assessment of computational thinking concepts, in: Proc. IEEE Symp. Vis. Lang. Hum.-Centric Comput. (VL/HCC), IEEE, 2016, pp. 238–239, https://doi.org/10.1109/VLHCC.2016.7739695.

[66] T. Palts, A model for assessing computational thinking skills. Doctoral Dissertation, University of Tartu, University of Tartu Press, Estonia, 2021.

[67] A. Piatti, G. Adorni, L. El-Hamamsy, L. Negrini, D. Assaf, L. Gambardella, F. Mondada, The CT-Cube: a framework for the design and assessment of computational thinking activities, Comput. Hum. Behav. Rep 5 (2022) 100166, https://doi.org/10.1016/j.chbr.2021.100166.

[68] E. Polat, S. Hopcan, S. Küçük, B. Şişman, A comprehensive assessment of secondary school students' computational thinking skills, Br. J. Educ. Technol. 52 (5) (2021) 1965–1980.

[69] Y.F. Putri, K. Kadir, A. Dimyati, Analysis of content validity on mathematical computational thinking skill test for junior high school students using aiken method, J. Hipotenusa 4 (2) (2022) 108–123, https://doi.org/10.18326/hipotenusa.v4i2.7465.

[70] E. Relkin, Assessing Young Children's Computational Thinking Abilities, Tufts University, 2018. Master's thesis.

[71] E. Relkin, M.U. Bers, Designing an assessment of computational thinking abilities for young children, in: Coding as a Playground, Routledge, New York, 2019, https://doi.org/10.4324/9780429453755-5.

[72] E. Relkin, M.U. Bers, TechCheck-K: a measure of computational thinking for kindergarten children, in: Proc. IEEE Glob. Eng. Educ. Conf. (EDUCON), IEEE, 2021, pp. 1696–1702, https://doi.org/10.1109/EDUCON46332.2021.9453926.

[73] E. Relkin, L.E. de Ruiter, M.U. Bers, TechCheck: development and validation of an unplugged assessment of computational thinking in early childhood education, J. Sci. Educ. Technol. 29 (2020) 482–498, https://doi.org/10.1007/s10956-020-09831-x.

[74] J.A. Rodríguez-Martínez, J.A. González-Calero, J.M. Sáez-López, Computational thinking and mathematics using scratch: an experiment with sixth-grade students, Interact. Learn. Environ. 28 (3) (2020) 316–327, https://doi.org/10.1080/10494820.2019.1612448.

[75] M. Román-González, J.C. Pérez-González, C. Jiménez-Fernández, Which cognitive abilities underlie computational thinking? Criterion validity of the computational thinking test, Comput. Hum. Behav. 72 (2017) 678–691, https://doi.org/10.1016/j.chb.2016.08.047.

[76] M. Román-González, J.C. Pérez-González, J. Moreno-León, G. Robles, Can computational talent be detected? Predictive validity of the computational thinking test, Int. J. Child-Comput. Interact. 18 (2018) 47–58, https://doi.org/10.1016/j.ijcci.2018.06.004.

[77] E. Rowe, M.V. Almeda, J. Asbell-Clarke, R. Scruggs, R. Baker, E. Bardar, S. Gasca, Assessing implicit computational thinking in zoombinis puzzle gameplay, Comput. Hum. Behav. 120 (2021) 106707, https://doi.org/10.1016/j.chb.2021.106707.

[78] E. Rowe, J. Asbell-Clarke, M.V. Almeda, S. Gasca, T. Edwards, E. Bardar, V. Shute, M. Ventura, Interactive assessments of CT (IACT): Digital interactive logic puzzles to assess computational thinking in grades 3–8, Int. J. Comput. Sci. Eng. Syst. 5 (2) (2021) 28–73, https://doi.org/10.21585/ijcses.v5i1.149.

[79] L.M. Seiter, B. Foreman, Modeling the learning progressions of computational thinking of primary grade students, in: Proc. 9th Int. Comput. Educ. Res. Conf. ACM, 2013, https://doi.org/10.1145/2493394.2493403.

[80] K. Sigayret, A. Tricot, N. Blanc, Unplugged or plugged-in programming learning: a comparative experimental study, Comput. Educ. 184 (2022) 104505, https://doi.org/10.1016/j.compedu.2022.104505.

[81] J. Sung, Assessing young Korean children's computational thinking: a validation study of two measurements, Educ. Inf. Technol. 27 (2022) 12969–12997, https://doi.org/10.1007/s10639-022-11137-x.

[82] Ž. Ternik, L. Todorovski, I. Nančovska Šerbec, Assessing the agreement in the bebras tasks categorisation, in: K. Kori, M. Laanpere (Eds.), Informatics in Schools: Engaging Learners in Computational Thinking. Lecture Notes in Computer Science, 12518, Springer, Cham, 2020, pp. 35–48, https://doi.org/10.1007/978-3-030-63212-0_3.

[83] Y. Tran, Computational thinking equity in elementary classrooms: what third-grade students know and can do, J. Educ. Comput. Res. 57 (1) (2019) 3–31, https://doi.org/10.1177/0735633117743918.

[84] M.J. Tsai, J.C. Liang, C.Y. Hsu, The computational thinking scale for computer literacy education, J. Educ. Comput. Res. 59 (4) (2021) 579–602, https://doi.org/10.1177/0735633120972356.

[85] M.J. Tsai, F.P. Chien, S.W.Y. Lee, C.Y. Hsu, J.C. Liang, Development and validation of the computational thinking test for elementary school students (CTT-ES): Correlate CT competency with CT disposition, J. Educ. Comput. Res. 60 (5) (2022) 1110–1129, https://doi.org/10.1177/07356331211051043.

[86] M.J. Tsai, J.C. Liang, S.W.Y. Lee, C.Y. Hsu, Structural validation for the developmental model of computational thinking, J. Educ. Comput. Res. 60 (1) (2022) 56–73, https://doi.org/10.1177/07356331211017794.

[87] L. Wang, F. Geng, X. Hao, et al., Measuring coding ability in young children: relations to computational thinking, creative thinking, and working memory, Curr. Psychol. 42 (2023) 8039–8050, https://doi.org/10.1007/s12144-021-02085-9.

[88] L. Werner, J. Denner, S. Campe, D.C. Kawamoto, The fairy performance assessment, in: Proc. 43rd ACM Tech. Symp. Comput. Sci. Educ. (SIGCSE), ACM, 2012, pp. 215–220, https://doi.org/10.1145/2157136.2157200.

[89] S.Y. Wu, Y.S. Su, Visual programming environments and computational thinking performance of fifth- and sixth-grade students, J. Educ. Comput. Res. 59 (6) (2021) 1075–1092, https://doi.org/10.1177/0735633120988807.

[90] M. Zapata-Cáceres, E. Martín-Barroso, M. Román-González, Computational thinking test for beginners: design and content validation, in: Proc. IEEE Glob. Eng. Educ. Conf. (EDUCON). IEEE, 2020, pp. 1905–1914, https://doi.org/10.1109/EDUCON45650.2020.9125368.

[91] Z. Zhan, W. He, X. Yi, S. Ma, Effect of unplugged programming teaching aids on children's computational thinking and classroom interaction: with respect to Piaget's four stages theory, J. Educ. Comput. Res. 60 (5) (2022) 1130–1155, https://doi.org/10.1177/07356331211057143.

[92] S. Zhang, G.K.W. Wong, G. Pan, Computational thinking test for lower primary students: design principles, content validation, and pilot testing, in: Proc. IEEE Int. Conf. Eng. Technol. Educ. (TALE). IEEE, 2021, pp. 345–352, https://doi.org/10.1109/TALE52509.2021.9678852.

[93] S. Zhang, G.K.W. Wong, Development and validation of a computational thinking test for lower primary school students, Educ. Technol. Res. Dev. 71 (4) (2023) 1595–1630, https://doi.org/10.1007/s11423-023-10231-2.

[94] W. Zhao, V.J. Shute, Can playing a video game foster computational thinking skills? Comput. Educ. 141 (2019) 103633 https://doi.org/10.1016/j.compedu.2019.103633.

[95] B. Zhong, Q. Wang, J. Chen, Y. Li, An exploration of three-dimensional integrated assessment for computational thinking, J. Educ. Comput. Res. 53 (4) (2016) 562–590, https://doi.org/10.1177/0735633115608444.

[96] C. Hederich-Martínez, Análisis Cuantitativo De Datos Para La Investigación Educativa Y Social, Universidad Pedagógica Nacional, Bogotá, 2022. http://hdl.handle.net/20.500.12209/18427.

[97] A. Blanco, Fiabilidad y generalización de la observación conductual, Anu. Psicol. 43 (1989) 43–58.

[98] R.J. de Ayala, The Theory and Practice of Item Response Theory, Guilford Press, New York, 2009.

[99] T.A. Warm, Weighted likelihood estimation of ability in item response theory, Psychometrika 54 (3) (1989) 427–450, https://doi.org/10.1007/BF02294627.

[100] M. Rubio-Stipec, H.R. Bird, G. Canino, M.S. Gould, The internal consistency of psychiatric symptoms scales: reliability is not population-specific, JCPP (J. Child Psychol. Psychiatry) 31 (5) (1990) 693–702.

[101] C.H. Yu, Test-retest reliability, in: K. Kempf-Leonard (Ed.), Encyclopedia of Social Measurement, vol. 3, Academic Press, San Diego, 2005, pp. 777–784.

[102] R.M. Kaplan, D.P. Saccuzzo, Psychological Testing: Principles, Applications, and Issues, Cengage Learning, Belmont, 2008.

[103] R.J. Cohen, M.E. Swerdlik, Psychological Testing and Assessment, ninth ed., McGraw-Hill Education, New York, 2018.

[104] M.L. McHugh, Interrater reliability: the kappa statistic, Biochem. Med. 22 (3) (2012) 276–282, https://doi.org/10.11613/BM.2012.031.

[105] J. Cerda, L. Villarroel, Evaluación de la concordancia interobservador en investigación pediátrica: coeficiente de Kappa, Rev. Chil. Pediatr. 79 (1) (2008) 54–58, https://doi.org/10.4067/S0370-41062008000100008.

[106] M.S.B. Yusoff, ABC of content validation and content validity index calculation, Educ. Méd. J. 11 (2) (2019) 49–54, https://doi.org/10.21315/eimj2019.11.2.6.

[107] G. Gilbert, S. Prion, Making sense of methods and measurement: lawshe's content validity index, Clin. Simul. Nurs. 12 (2016) 530–531, https://doi.org/10.1016/j.ecns.2016.08.002.

[108] K. Coaley, An Introduction to Psychological Assessment and Psychometrics, second ed., Sage Publications, London, 2010.

[109] W.L. Lin, G. Yao, Concurrent validity, in: A.C. Michalos (Ed.), Encyclopedia of Quality of Life and Well-Being Research, Springer, Dordrecht, 2014, https://doi.org/10.1007/978-94-007-0753-5_516.

[110] M. Alavi, E. Biros, M. Cleary, Notes to factor analysis techniques for construct validity, Can. J. Nurs. Res. 56 (2) (2024) 164–170, https://doi.org/10.1177/08445621231204296.

[111] A. Anastasi, S. Urbina, Psychological Testing, seventh ed., Prentice Hall, Upper Saddle River, 1997.

[112] R.F. Zuñiga, J.A. Hurtado, G. Robles, Evaluación de las habilidades de pensamiento computacional: Revisión sistemática de la literatura, Rev. Iberoam. Tecnol. Aprendiz. 18 (4) (2023).

[113] H.C. Şenel, S. Şenel, Assessment of computational thinking skills, in: M. Saritepeci, H. Yildiz Durak (Eds.), Integrating Computational Thinking Through Design-based Learning, Springer, Singapore, 2024, https://doi.org/10.1007/978-981-96-0853-9_12.

[114] M. Yağcı, A valid and reliable tool for examining computational thinking skills, Educ. Inf. Technol. 24 (2019) 929–951, https://doi.org/10.1007/s10639-018-9801-8.

[115] J. Waite, Pedagogy in Teaching Computer Science in Schools: a Literature Review, The Royal Society, London, 2017. https://royalsociety.org/-/media/policy/projects/computing-education/literature-review-pedagogy-in-teaching.pdf.

[116] Ö. Korkmaz, X. Bai, Adapting computational thinking scale (CTS) for Chinese high school students and their computational thinking skills level, Particip. Educ. Res. 6 (1) (2019) 10–26, https://doi.org/10.17275/PER.19.2.6.1.

[117] L. Zhao, X. Liu, C. Wang, Y.S. Su, Effect of different mind mapping approaches on primary school students' computational thinking skills during visual programming learning, Comput. Educ. 181 (2022) 104445, https://doi.org/10.1016/j.compedu.2022.104445.

[118] A. Basawapatna, H. Kyu, K.H. Koh, A. Repenning, D. Webb, K. Marshall, Recognizing computational thinking patterns, in: Proc. 42nd ACM Tech. Symp. Comput. Sci. Educ. (SIGCSE), ACM, 2011, pp. 245–250, https://doi.org/10.1145/1953163.1953241.

[119] J. Denner, L. Werner, E. Ortiz, Computer games created by middle school girls: can they be used to measure understanding of computer science concepts? Comput. Educ. 58 (1) (2012) 240–249, https://doi.org/10.1016/j.compedu.2011.08.006.
[120] K.H. Koh, A. Basawapatna, H. Nickerson, A. Repenning, Real-time assessment of computational thinking, in: Proc. IEEE Symp. Vis. Lang. Hum.-Centric Comput. (VL/HCC), IEEE, 2014, pp. 49–52, https://doi.org/10.1109/VLHCC.2014.6883021.
[121] V. Dagienė, S. Sentance, It's computational thinking! bebras tasks in the curriculum, , in: A. Brodnik, F. Tort (Eds.), Informatics in Schools: Improvement of Informatics Knowledge and Perception, Lecture Notes in Computer Science, 9973, , Springer, Cham, 2016, pp. 28–39, https://doi.org/10.1007/978-3-319-46747-4_3.