EUROPEAN JOURNAL OF STEM EDUCATION

European Journal of STEM Education,
2024, 9(1), 02
ISSN: 2468-4368

lectito

Review Article

A Systematic Umbrella Review on Computational Thinking Assessment in Higher Education

Xiaoling Zhang [1]*, Fenia Aivaloglou [1], Marcus Specht [1]

[1] *Delft University of Technology*, NETHERLANDS

***Corresponding Author:** x.zhang-14@tudelft.nl

Citation: Zhang, X., Aivaloglou, F. and Specht, M. (2024). A Systematic Umbrella Review on Computational Thinking Assessment in Higher Education. *European Journal of STEM Education*, 9(1), 02. https://doi.org/10.20897/ejsteme/14175

Published: January 31, 2024

ABSTRACT

Computational Thinking (CT) is considered a core 21st century digital skill. The aspect of assessment is crucial and knowing what, who, when, how, and where to assess is important for assessment design. In this study, we conducted an umbrella review to gain insights regarding CT assessment in higher education. In total, we analyzed 11 reviews, focusing on: (1) bibliographical and methodological characteristics of the reviews; (2) aspects relevant of assessment design, including a) assessed constructs, b) applied assessment methodologies, and c) assessment contexts. Our findings suggest an increased attention on this topic. However, hardly any reviews reasoned the selection of their review methodology, and most of the reviews did not thoroughly examine existing reviews. Regarding assessment design aspects, most reviews did not confine their scope to higher education; however, findings on interventions and educational settings show commonalities. We identified 120 unique assessed constructs and around 10 types of assessment methods. Though a combined use of distinct assessment methods is suggested in reviews, guidelines for appropriate assessment design are yet to be constructed. Based on the findings, we argue that it is necessary to explore different combinations of assessment design in various contexts to construct assessment guidelines.

Keywords: umbrella review, computational thinking, assessment, higher education

INTRODUCTION

Computational Thinking (CT) is currently regarded as a core digital competency and it has been widely promoted in K-12 education as well as higher education settings. Wing (2011) defined CT as a set of problem-solving skills with which people can formulate the solution for problems such that the solutions can be carried out by any information processing agent. More recently, CT is stated a thinking model that human beings utilized for problem-solving regardless of the rapid change of technology throughout history (Denning and Tedre, 2019). Existing studies investigated not only the definition of CT, but also topics on various aspects of CT teaching and learning (Lyon and Magana, 2020; Tang et al., 2020; Cutumisu et al., 2019). One can find a consensus on the importance of CT and various organizations have stressed that in the last decade (Lyon and Magana, 2020; Tang et al., 2020; Cutumisu et al., 2019); however, more effort is needed to explore further what CT is, how CT can be properly operationalized in educational activities, what distinguishes CT from other kinds of thinking skills, and how CT can be incorporated within other subject domains. While the emphasis of CT education was more on the

K-12 educational settings in early days, it has gradually appealed more and more attention in higher education. In this paper, CT is deemed as a skill independent of the conventional programming or computing skills.

Assessment can be a means to assure the quality of education and to support students in their learning process. Examining how a competency is assessed can help us understand the operationalization and the incorporation of the competency. A literature review is an intuitive method for gaining a holistic review on CT assessment in higher education. Though several existing reviews examined CT assessment from different aspects, as far as we are aware, an umbrella review which revises and summarizes knowledge accumulated on CT assessment in higher education is yet to be presented.

In this paper, we synthesize existing knowledge by performing a systematic umbrella review on reviews relevant to CT assessment in higher education. Our research questions are as follows:

- RQ1 What are the features of the existing reviews? Sub-questions are listed as follows:
  - RQ1a What are the bibliographical characteristics of the reviews, such as year of publication, country of the work, type of publication?
  - RQ1b What are the methodological characteristics of the reviews, such as type of review, principles, methodology followed for the review and tools used for reviewing?
- RQ2 What are the explored aspects of CT assessment in the reviews? Sub-questions are listed as follows:
  - RQ2a Which constructs have been found in CT assessment?
  - RQ2b Which aspects characterize the methodology applied for CT assessment?
  - RQ2c What characterizes the contexts in which CT has been assessed?

Preliminary results of this work were presented by Zhang and Specht (2022). According to our preliminary results, existing reviews hardly refer to other reviews and rarely synthesize existing knowledge which can provide the readers a comprehensive view of the field. Therefore, in this work, by investigating all existing reviews (as exhaustive as it can be), we aim to gather and synthesize knowledge to picture the field. In this paper, we extend the previous work with (1) an extensive analysis of the publication features, with discussion on the types of publication and the trends that emerge in the publications (RQ1a), (2) an extensive analysis of the methods adopted by the included reviews (RQ1b), (3) an overview of the aspects investigated on the topic of CT assessment in higher education (RQ2), and (4) a comparison and analysis of the following topics: the assessment constructs (RQ2a), the assessment methodology (RQ2b), and the assessment context (RQ2c).

RELATED WORK

CT and Its Assessment in Higher Education

A considerable amount of knowledge has been accumulated on the topic of CT education. Theoretical frameworks have been established over the years to facilitate CT understanding and promotion, such as Brennan and Resnick's three-dimensional framework (Brennan and Resnick, 2012), Weintrop's taxonomy of CT in mathematics and science (Weintrop et al., 2016), and Grover and Pea's competency framework (Grover et al., 2017), which are seemingly the most frequently adopted in the literature. Tools such as Alice, Scratch, and Bebras, have been developed for teaching, learning and assessment of CT in different contexts (Cutumisu et al., 2019). Moreover, curricula have also been developed for teaching CT in different contexts (Hsu et al., 2018).

It is noteworthy that CT has been mostly coupled with CS and programming, while it is regarded as a skill which can be applied to not only domains such as engineering and mathematics, but also daily life scenarios (Angeli and Giannakos, 2020; Li and Lefevre, 2020; Tedre and Denning, 2016; Wing, 2016). Moreover, contributions focused more in the context of K-12 than in higher education (Cutumisu et al., 2019). It is only after 2014 that research on CT in higher education started to develop briskly, with a focus on operationalizing and incorporating CT in programming and CS contexts (Cutumisu et al., 2019). Meanwhile, assessment of CT is highly diverse due to the different implementations and operationalizations of CT in different educational contexts (Selby and Woollard, 2014). Furthermore, the assessments are rarely validated and most of the work suggests a combination use of different assessment methods (Lyon and Magana, 2020; Cutumisu et al., 2019; Li and Lefevre, 2020; Ezeamuzie and Leung, 2022; Lu et al., 2022).

Reviews on CT Assessment in Higher Education

According to the preliminary results (Zhang and Specht, 2022), there are 11 reviews that either fully focus on CT assessment in higher education or discuss CT assessment in higher education as one of their components. Among the reviews that mainly focus on the topic of CT assessment, de Araujo et al. (2016) investigated specifically the assessed abilities, Haseski and Ilic (2019) focused on analyzing features of paper-and-pencil data collection instruments, Vinu Varghese and Renumol (2021) and Poulakis and Politis (2021) examined the topic by studying methods and approaches appeared in the studies while Tang et al. (2020) looked more holistically how CT has

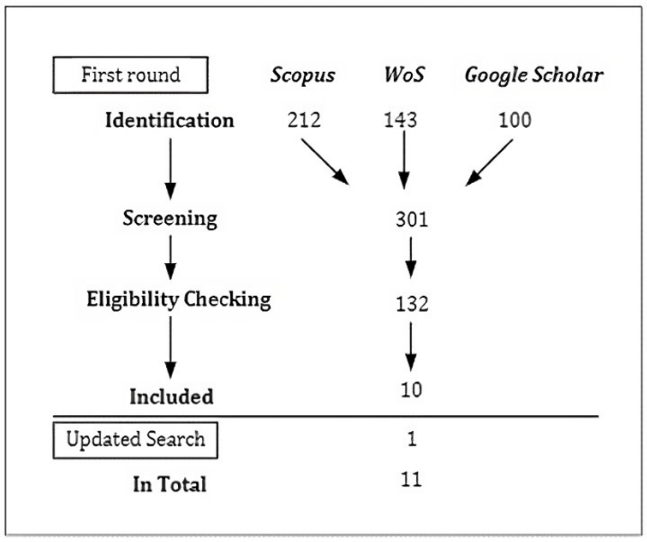

Figure 1. PRISMA flowchart with results

Table 1. Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
• The type of the study must be a review. • The study contains content about assessment of CT in higher education.	• The study is inaccessible (the full-text article is neither available in the databases or upon the request to the authors). • The study is not peer-reviewed. • The study is written in non-English language. • The study is a duplicate of the existing ones.

been assessed in studies, and Cutumisu et al. (2019) and Lu et al. (2022) analyzed vital features of CT assessed, with the work of Lu et al. (2022) specifically focusing on higher education. For other reviews, Lyon and Magana (2020) and de Jong and Jeuring (2020) described CT assessment methods as part of the results for investigation of key areas for future study and interventions used in higher education, respectively. Meanwhile, Taslibeyaz et al. (2020) investigated both conceptual understanding and measurement approaches for CT development and Ezeamuzie and Leung (2022) studied various operationalization of CT in the literature, presenting CT assessment as part of their results.

METHOD

An umbrella review is a review of reviews on a topic that compiles all the evidence of existing reviews to present a high-level overview (Fusar-Poli and Radua, 2018). With more studies emerging on the topic of CT assessment, it is essential for researchers to have an overview of the field; to the best of our knowledge, there is not a study which provides a high-level overview. To address this gap and answer the research questions, we conduct a systematic umbrella review to investigate the topic of CT assessment in higher education.

This study adopts a systematic process for gathering data with the following steps (Taylor, 2012): (1) identify the scope and formulate research questions, (2) plan the review and create a document protocol, (3) develop inclusion and exclusion criteria, (4) search and screen the studies, (5) extract and synthesize data. We conducted a narrative analysis on literature to narrow the scope of our work and establish a document protocol. Inclusion and exclusion criteria were developed with expert's suggestions. The PRISMA flowchart was adopted for recording the results of the first four steps, while all other data were documented in an Excel file. Aspects of the included reviews were examined through discussions between the authors where necessary. More details about the whole process is presented in the following subsections. The principal results for the key steps are shown in **Figure 1**.

Inclusion and Exclusion Criteria

We developed four exclusion criteria and two inclusion criteria to identify the eligible reviews for further analysis. The criteria are listed in **Table 1**.

Table 2. Bibliographical features of the reviews on CT assessment in higher education

Characteristics	Results and frequency
Publication year	2016 (1), 2019 (2), 2020 (4), 2021 (3), 2022 (1)
Region of the first author	Brazil (1), Canada (2), Greece (1), Hongkong (1), India (1), The Netherlands (1), the United States (2), Turkey (2)
Publication outlet	Journal (7), Conference (3), Book chapter (1)

Identification and Filtering the Reviews

Data sources and search query

Two world-leading citation databases (Zhu and Liu, 2020) were used as main data sources to ensure the coverage of the work: Web of Science (WoS) and Elsevier journal database (SCOPUS). Google Scholar was used as an additional source to identify relevant records. We used "computational thinking" and "review" as keywords for search queries; "higher education" was left out to include all potentially relevant records. Search queries are listed as follows (first round search at 2021-08-25):

- Web of Science: TS = ("computational thinking" AND "review")
- SCOPUS: TITLE-ABS-KEY ("computational thinking" AND "review")

With these queries, we retrieved records in Scopus and Web of Science that contain "computational thinking" and "review" in at least one of the following fields: abstract, title, and key words. As for Google Scholar, we examined the first 10 pages of Google Scholar results, sorted by relevancy, with the search query "computational thinking" and "review". In total, we identified 355 records from SCOPUS and WoS and the first 10 pages (100 records) of Google Scholar results for screening and eligibility checking.

Screening and eligibility checking

For all records identified from WoS, SCOPUS and Google Scholar, we applied the inclusion and exclusion rules to their title, abstract and keywords for preliminary filtering of the records. Before the screening stage, we removed duplicated records, resulting in 301 records. At the screening stage, by reading titles, key words, and abstracts, we removed records that were non-English, not peer-reviewed, or inaccessible, as well as records that did not discuss CT assessment in higher education or were not reviews. This resulted in 132 records for eligibility checking. For eligibility checking, we checked the full texts to exclude the records that were not reviews or did not discuss the assessment of CT in higher education, leaving 10 reviews for data extraction and synthesis. Furthermore, an updated search on 2022-02-02 in all three databases with the same procedure resulted in one more eligible review. In total, we included a total number of 11 reviews (Denning and Tedre, 2019; Lyon and Magana, 2020; Tang et al., 2020; Selby and Woollard, 2014; Ezeamuzie and Leung, 2022; Lu et al., 2022; de Araujo et al., 2016; Haseski and Ilic, 2019; Vinu Varghese and Renumol, 2021; Poulakis and Politis, 2021; de Jong and Jeuring, 2020) for analysis in this work.

Data Extraction and Synthesis

To answer our research questions, we qualitatively analyzed the included reviews with multiple features using Excel and ATLAS.ti as the data extraction and analysis tools. To answer RQ1, we examined the meta-information of the reviews and the title, abstract and keywords of the reviews to extract bibliographical characteristics. As for the methodological characteristics of the reviews, we examined the abstract and methodological description section. Then, to answer RQ2, we analyzed the reviews with an emphasis on abstract, introduction, method, and result sections. The analysis of the information extracted for research questions RQ1a and RQ2a to RQ2c followed a bottom-up approach; we first identified and marked all relevant information and then aggregated and synthesized those information according to the research questions.

RESULTS AND ANALYSIS

RQ1 Bibliographical and Methodological Characteristics of the Reviews

RQ1a – Bibliographical characteristics of the reviews

We extracted the following features of the reviews from the meta-information as bibliographic characteristics: year of publication, region of the work, and publication outlet. The results are shown in **Table 2**.

The results show that most reviews were published within the last five years and no reviews were found before 2016. This increase in the number of reviews indicates a growing interest in the topic. As we look at the region of

Table 3. Methodological characteristics of the examined studies on CT assessment in higher education

Characteristics	Results and frequency
Type of review	Systematic review (5) Scoping review (3) Narrative review (2) Systematic mapping study (1)
Principles and methodologies	Adopted existing methods or guidelines (10) No specification on methods used (1)
Searched databases	ACM / ACM Digital Library (6) EBSCO (1) ERIC (9) Engineering Village: Inspec (1) Engineering Village: Compendex (1) Google Scholar (2) IEEE Xplore (5) PsycINFO (1) ProQuest (1) ScienceDirect (5) Scopus (6) Springer / SpringerLink (5) Web of Science (3)
Selection criteria	Included structured description of inclusion criteria and exclusion criteria (9) Described some criteria for selection of the studies (2)
Quality of the review process	Described how discrepancies were addressed (8) Not specified (3)
Scope of the review	CT assessment in higher education as: • Dominant topic (1) • Non-dominant topic (10)
	Type of evidence included: • Empirical studies (7) • Not-specified (4)

the reviews, we can assert that the contribution to this research topic is international, with the United States, Turkey and Canada being slightly more active. Meanwhile, the contributions were published as journal articles, conference papers, and book chapters in journals such as Informatics in Education, conferences such as Frontiers in Education Conference and books such as Research on E-Learning and ICT in Education.

RQ1b – Methodological characteristics of the reviews

For all reviews, we identified the following methodological characteristics: type of review, principles and methodology adopted, searched databases, inclusion and exclusion criteria (e.g. education level), quality control, and focus of review. The results are shown in **Table 3**.

To start with, for the *type of reviews*, we found five systematic reviews while the others were scoping reviews, narrative reviews, and a systematic mapping study. Regarding the *principles and methodologies* applied in the reviews, even though 10 out of 11 reviews followed or referred to existing methods or guidelines, hardly any review thoroughly explains the choice of the adopted methods. In terms of the *searched databases*, every review used two to six databases as data sources, and there were in total 15 databases identified as data sources, being either domain-specific databases or general indexing databases. Among the domain-specific databases, ERIC was the most used database, followed by the general indexing database Scopus. Moreover, the eligibility of the retrieved items was checked against *selection criteria* in all reviews, with several of the reviews specifying both the inclusion and exclusion criteria. Regarding the *review process*, most of the reviews described how discrepancies in the selection of reviewed studies and the collection and analysis of information were resolved, while the rest provided no indication on that topic.

Last but not the least, while sharing some similarities, the *scope of the reviews* differ from each other, meaning that each review focuses on different aspects. Among all reviews, only Lu et al. (2022) exactly scoped their review on the topic of CT assessment in higher education by examining empirical evidence. For the reviews which limited their scope to higher education, one provided a general picture of CT in higher education (Lyon and Magana, 2020) and another (de Jong and Jeuring, 2020) studied CT interventions in higher education from empirical experiences, with assessment of CT being only a part of the investigation. For the rest of the other reviews with which the scope was not confined to higher education context, they investigated assessment of CT across all educational levels including: assessment of CT (de Araujo et al., 2016; Poulakis and Politis, 2021), data collection instruments

Table 4. Aspects investigated on CT assessment in higher education

Aspects of CT assessment	References
Assessment constructs (8)	Tang et al. (2020); Cutumisu et al. (2019); Lu et al. (2022); de Araujo et al. (2016); Haseski and Ilic (2019); Poulakis and Politis (2021); de Jong and Jeuring (2020); Taslibeyaz et al. (2020)
Assessment methodologies (10)	Lyon and Magana (2020); Tang et al. (2020); Cutumisu et al. (2019); Ezeamuzie and Leung (2022); Lu et al. (2022); de Araujo et al. (2016); Haseski and Ilic (2019); Vinu Varghese and Renumol (2021); Poulakis and Politis (2021); Taslibeyaz et al. (2020)
Assessment contexts (11)	Taylor (2012); Lyon and Magana (2020); Tang et al. (2020); Selby and Woollard (2014); Ezeamuzie and Leung (2022); Lu et al. (2022); de Araujo et al. (2016); Haseski and Ilic (2019); Vinu Varghese and Renumol (2021); Poulakis and Politis (2021); de Jong and Jeuring (2020)

Table 5. An overview of the methods and analysis applied for the investigation of the assessed constructs

Method	References
Apply an existing classification scheme (1)	Tang et al. (2020)
Utilize an existing framework (2)	Cutumisu et al. (2019); Lu et al. (2022)
Natural clustering of constructs emerged from the study (5)	de Araujo et al. (2016); Haseski and Ilic (2019); Poulakis and Politis (2021); de Jong and Jeuring (2020); Taslibeyaz et al. (2020)

for CT assessment (Haseski and Ilic, 2019), empirical experiences of CT assessments (Tang et al., 2020; Cutumisu et al., 2019), empirical experiences on development of CT (Taslibeyaz et al., 2020), empirical experiences of CT practices (Ezeamuzie and Leung, 2022), and assessment methods and interventions for developing CT (Vinu Varghese and Renumol, 2021). Additionally, seven out of 11 reviews investigated empirical evidence which collected information through observation and documentation of certain behavior and patterns or through an experiment.

RQ2 – Overview of Aspects of CT Assessment Examined in the Reviews

Table 4 presents an overview of aspects investigated on CT assessment in higher education with regard to the RQ2 and its sub-questions.

Regarding *assessment constructs*, a total of eight reviews examined assessed constructs and their characteristics regarding the skills, and competencies. Of the other three reviews, two examined the definition of CT with which the assessed constructs can be deduced according to the constructive alignment theory (Biggs, 1996). As for *assessment methodologies*, except for the work from de Jong and Jeuring (2020), all reviews examined perspectives relevant to the delivery of the assessment, namely instruments and tools developed for assessment and their characteristics, categorization of assessment methods, and quality indicators for the assessment methods. In terms of *assessment contexts*, all reviews investigated this topic with different focus on the following perspectives: academic domain, education level, educational setting, and intervention. Details are presented in the following subsections.

RQ2a – Characteristics of the assessed constructs

Table 4 shows that eight out of 11 reviews examined the assessed constructs. Regarding the method applied for studying assessed constructs, as presented in **Table 5**, by applying (a) Sullivan and Heffernan (2016)'s categories of first-order and second-order learning outcomes regarding the relationship between CT and other subject domains and (b) McMillan (2013)'s categories of cognitive constructs and non-cognitive dispositions and skills, Tang et al. (2020) examined the relationship between the assessed constructs and the subject domain. For the remaining set of reviews, Cutumisu et al. (2019) and Lu et al. (2022) utilized an existing CT framework as reference to identify the assessed constructs, while the others clustered and synthesized similar constructs emerged in their investigation (de Araujo et al., 2016; Haseski and Ilic, 2019; Poulakis and Politis, 2021; de Jong and Jeuring, 2020; Taslibeyaz et al., 2020).

As for assessed constructs and their categorization, Cutumisu et al. (2019) and Lu et al. (2022) mapped the assessed constructs to Brennan and Resnick's (2012) three-dimensional framework and a hybrid framework inferred from three frameworks (Brennan and Resnick, 2012; Weintrop et al., 2016; Grover et al., 2017), respectively. Though both frameworks sketched CT competencies as a three-dimensional framework, including CT concepts, practices and perspectives, the hybrid one was claimed to be more generic and independent from subjects that allows a broader coverage of CT skills. The other five reviews presented: (1) only the overarching categories of assessed constructs that provide a high-level categorization of assessed constructs (de Jong and Jeuring, 2020; Taslibeyaz et al., 2020); (2) only the assessed constructs (de Araujo et al., 2016; Poulakis and Politis, 2021); (3) both the constructs and the overarching categories (Haseski, H. I. and Ilic, 2019). Taslibeyaz et al. (2020) and de Jong and Jeuring (2020) identified a total number of six distinct categories, including attitude towards CT, attitude-motivation, CT knowledge, CT skills, problem-solving skills, and programming skills without specification

Table 6. CT assessment constructs appearing at least three times in the included reviews

Category	Constructs & frequency	Definition
CT concepts	Algorithm / algorithmic thinking / algorithm skills (5)	The skills involved in developing an algorithm which is precise with step-by-step procedures to solve a problem (Grover et al., 2017).
	Data / data analysis / data collection, data analysis / data representation (5)	Including storing, retrieving, updating values as well as analyzing, and visualizing data (Brennan and Resnick, 2012; Weintrop et al., 2016).
	Automation / automating solutions (4)	A key component of computational thinking that aims at a solution to be executed by a machine (Grover et al., 2017).
	Logic / logic and logical thinking (4)	Analyzing situations to make a decision or reach a conclusion about a situation (Grover et al., 2017).
	Critical thinking (3)	Not found.
	Evaluation (3)	Solutions to problems are evaluated for accuracy and correctness with respect to the desired result or goal (Grover et al., 2017).
	Pattern / pattern recognition (3)	Pattern recognition in CT could result in a definition of a generalizable solution which can utilize automation in computing for dealing with a generic situation (Grover et al., 2017).
	Synchronization / synchronize (3)	Not found.
CT practices	Abstraction (5)	Abstraction is 'information hiding'. Through 'black-box'- ing details, one can focus only on the input and output and provides a way of simplifying and managing complexity (Weintrop et al., 2016; Grover et al., 2017).
	Problem-solving (4)	Not found.
	Modularity / modularizing / modelling (3)	Building something large by putting together collections of smaller parts is an important practice for all design and problem solving (Brennan and Resnick, 2012).
	Testing / testing and debugging (3)	Practices that are relevant to dealing with – and anticipating – problems include identifying the issue, systematically evaluating the system to isolate the error and reproducing the problem so that potential solutions can be tested reliably (Weintrop et al., 2016; Grover et al., 2017).
CT perspectives	Creativity and creation (4)	Creativity as a CT practice acts on two levels – it aims to encourage out-of-the-box thinking and alternative approaches to solving problems; and it aims to encourage the creation of computational artefacts as a form of creative expression (Grover et al., 2017).
	Collaboration and cooperation (3)	Perspectives about working with CT skills in a collaborative or cooperative format (Grover et al., 2017).

on the differences of categories. Meanwhile, de Araujo et al. (2016), Haseski and Ilic (2019), and Poulakis and Politis (2021) identified five categories, namely affective achievements towards CT, cognitive achievements towards CT, CT skills / abilities, CT concepts, CT patterns, and enumerated the underlying constructs.

It should be noted that some constructs were classified into distinct categories in different reviews as schemes used for classification varied from one review to another. Moreover, some categories nearly shared the same name, such as CT skills versus CT skills / ability, and attitudes towards CT versus attitude-motivation. However, in this work, we did not merge them since information on their meaning is insufficient at the current level of investigation. To sum up, our approach involved merging constructs or categories deemed identical, either because they align with the same definition or are presented in an identical manner.

With the constructs considered the same being merged, we identified a total of 120 unique assessed constructs. As they are too many to enumerate, by adopting the hybrid framework used in Lu et al. (2022)'s work, we present the assessed constructs which appeared in at least 3 reviews in **Table 6**. Example definitions of these constructs are provided accordingly.

RQ2b – Characteristics of the assessment methodology

As **Table 4** shows, 10 out of the 11 reviews studied different characteristics of the assessment methodology, mainly being *assessment methods*, *assessment tools*, *assessment instruments*, and *assessment quality indicators*.

As presented in **Table 7**, four reviews presented classifications of *assessment methods*. Lu et al. (2022)'s categorization resembles to the one used by Cutumisu et al. (2019) since it further investigated topics from Cutumisu et al. (2019)'s review; however, it focused more on higher education. Both of these two reviews identified the following types of assessment:

- block-based assessments – solving programming problems regardless of syntax by using programming blocks in block-based programming environments such as Scratch;

Table 7. Type of assessment methods identified from the reviews

Assessment methods	Reference
Block-based assessments	Cutumisu et al. (2019); Lu et al. (2022)
Knowledge / CT skill written tests / skill tests	Lyon and Magana (2020); Cutumisu et al. (2019); Lu et al. (2022)
Self-reported scales/survey	Lyon and Magana (2020); Cutumisu et al. (2019); Lu et al. (2022)
Robotics / game-based assessments (tangible tasks)	Cutumisu et al. (2019)
Combinations / using multiple forms of assessment	Lyon and Magana (2020); Cutumisu et al. (2019); Lu et al. (2022); Poulakis and Politis (2021)
Text-based programming assessment	Lu et al. (2022)
Course academic achievements of CS courses	Lu et al. (2022)
Interviews or interviews and observations	Lyon and Magana (2020); Lu et al. (2022)
Assignment / course grades	Lyon and Magana (2020)
Artefacts (classroom / students)	Lyon and Magana (2020)
Problems external to class	Lyon and Magana (2020)
Using specific programming environments	Poulakis and Politis (2021)
Using CT assessment criteria / psychometric tools	Poulakis and Politis (2021)

- CT skill written tests – using generic forms such as constructed response questions or multiple-choice questions to assess CT skills, e.g. Computational Thinking Knowledge test (CT Knowledge test);
- self-reported scales / surveys – mostly concerned with assessment of CT perspectives which includes inter- and intra-personal skills such as communication, collaboration, or questioning;
- using multiple forms of assessment – combining different types of assessment to gain a more holistic understanding of student understanding or mastery of skills, for example, the combined use of the CT Knowledge test and the block-based assessment in Scratch.

Additionally, Cutumisu et al. (2019) identified robotics/game-based assessments as a unique category to refer to assessments based on robotic tangible tasks or artefacts produced in game-based assessments such as AgentSheets. Meanwhile, Lu et al. (2022) identified three more categories compared to Cutumisu et al. (2019)'s work:

- text-based programming assessments – using text-based programming tasks to assess students' CT competency, for example, a Python programming task;
- interviews and observations – commonly used for studying practices of incorporating CT into traditional classrooms;
- course academic achievement – academic performance in coursework including students' achievement in quizzes, exams, projects and assignments.

Besides that, Lyon and J. Magana (2020)'s categorization resembled Lu et al. (2022) and Cutumisu et al. (2019)'s reviews, while being more generic. We noticed that, while these three reviews categorized assessment methods based on the type of assessment instrument and the medium used for assessment, Poulakis and Politis (2021) classified the assessment methods based on the characteristics of the medium. However, none of these classification schemes was constructed on a theory basis, but was rather established by summarizing the patterns that emerged in their studied sets of papers.

Additionally, some reviews presented additional aspects of assessment methodologies of CT, including *assessment tools* and *assessment instruments* used for the operationalization and delivery of assessment, and *assessment quality indicators* for controlling the quality of the assessment method. **Table 8** shows the reviews that investigated those topics, following the original categorization naming presented in the reviews. It is worth mentioning that no reviews explicitly describe what differentiated instruments from tools.

Regarding *assessment tools*, Lu et al. (2022) and Cutumisu et al. (2019) identified and reported them in a table, while the other reviews either examine the characteristics of the tools or categorize the tools according to certain characteristics and exemplify tools of those categories in text. Characteristics of the tools being studied include: the programming language necessary for the tools (Cutumisu et al., 2019), the type of the tool (Tang et al., 2020; Haseski and Ilic, 2019; Taslibeyaz et al., 2020), the functionality of the tool being either summative or formative (Taslibeyaz et al., 2020).

As for *assessment instruments*, characteristics identified include: the delivery format being either automatic or non-automatic (Cutumisu et al., 2019; Lu et al., 2022), and computer-based or paper-based (Cutumisu et al., 2019; Lu et al., 2022), the type of skills included being cognitive or non-cognitive (Cutumisu et al., 2019; Lu et al., 2022), the number of items and constructs included in the instrument (Haseski and Ilic, 2019), the type of assessment tasks in the instrument such as coding projects (Cutumisu et al., 2019).

Table 8. Additional characteristics of the assessment methodologies identified from the reviews

Aspects	Description	Reference
Assessment tools (6)	Aspects including the name of the tool, the characteristics of the tools, the programming language, the type of assessment tool, the functionality of the tool.	Tang et al. (2020); Cutumisu et al. (2019); Lu et al. (2022); Haseski and Ilic (2019); Vinu Varghese and Renumol (2021); Taslibeyaz et al. (2020)
Assessment instruments (5)	Aspects including the type of instruments and the characteristics of instruments, e.g. delivery format, type of skills included, type of instruments, number of factors, and items included in the instrument.	Cutumisu et al. (2019); Ezeamuzie and Leung (2022); Lu et al. (2022); de Araujo et al. (2016); Haseski and Ilic (2019)
Assessment quality indicators (2)	Validity and reliability indicators for the assessment methods.	Cutumisu et al. (2019); Haseski and Ilic (2019)

Table 9. Additional characteristics of the assessment methodologies identified from the reviews

Aspects	Description	Reference
Academic domain (7)	Concerned with the academic disciplines, program of studies, or subject matter for the assessed group of users.	Lyon and Magana (2020); Tang et al. (2020); Cutumisu et al. (2019); Ezeamuzie and Leung (2022); Lu et al. (2022); de Jong and Jeuring (2020); Taslibeyaz et al. (2020)
Education level (11)	Concerned with the level of education for the assessed group of users.	Lyon and Magana (2020); Tang et al. (2020); Cutumisu et al. (2019); Ezeamuzie and Leung (2022); Lu et al. (2022); de Araujo et al. (2016); Haseski and Ilic (2019); Vinu Varghese and Renumol (2021); Poulakis and Politis (2021); de Jong and Jeuring (2020); Taslibeyaz et al. (2020)
Educational setting (5)	Concerned with the type of educational activities that the assessed group of users were involved in.	Tang et al. (2020); Cutumisu et al. (2019); Ezeamuzie and Leung (2022); Lu et al. (2022); de Araujo et al. (2016)
Intervention (6)	Concerned with the actions taken for the development of skills and / or their corresponding characteristics.	Tang et al. (2020); Cutumisu et al. (2019); Ezeamuzie and Leung (2022); Vinu Varghese and Renumol (2021); de Jong and Jeuring (2020); Taslibeyaz et al. (2020)

Moreover, two reviews systematically grouped the *assessment quality indicators* for evaluating the quality of the assessment methods regarding the validity, the reliability, or both the validity and the reliability (Cutumisu et al., 2019; Haseski and Ilic, 2019).

RQ2c – Characteristics of the assessment context

As shown in **Table 4**, all reviews discussed assessment contexts. However, only a few examined every aspect of assessment context. **Table 9** presents the major aspects, being *academic domain*, *education level*, *educational setting*, and *intervention* that can affect the quality and the design of the assessment.

In total, seven reviews reported the *academic domain* of the assessment receiver, with different levels of detail; meanwhile, the others mentioned it as supplementary information, or presented it in an unstructured and uncomprehensive manner. de Jong and Jeuring (2020) presented a list of academic disciplines, while the others presented either information about the coursework within a study program, or a categorization of the coursework / the study program. Though there are various schemes for the categorizing the academic domain or the coursework, the reviews mainly distinguished the academic background according to the relevance of the study program to CS, Science, Technology, Engineering and Technology (Tang et al., 2020; Ezeamuzie and Leung, 2022; de Jong and Jeuring, 2020; Taslibeyaz et al., 2020). Furthermore, all reviews revealed that most of the contributions on CT were within the scope of STEM disciplines, CS and Programming Education, with an increase in the number of studies regarding Non-STEM and non-CS majors and subjects in recent years. This finding complies with the conclusions in several of the included reviews.

Regarding the *education level*, all reviews included it, with some exhaustively listed all grade levels that clearly indicates the year of study, whereas the others presented it at a higher level without detailed information on the year of study. For reviews that focused on higher education (Lyon and Magana, 2020; Lu et al., 2022; de Jong and Jeuring, 2020), Lu et al. (2022) grouped and tabulated the grade level according to the year of study, while the other two reviews regarded undergraduate as a category that covers all undergraduate years. The rest of the reviews that did not limit the scope to higher education, most of them made the distinction between the K-12 and undergraduate context, with some providing finer categorizations. Some reviews analyzed the distribution of

education levels of the included studies to gain insights into research trends (de Araujo et al., 2016; Haseski and Ilic, 2019), while some others such as Taslibeyaz et al. (2020) examined the tools or assessment methods used in certain levels of education.

Some reviews illustrated the assessment context with more details, including the *educational settings* and the *interventions*. Educational settings provide general information about the characteristics and the form of educational activities being formal or informal, representing in-school activities such as lectures or workshops, and after-school activities such as coding clubs. As shown in **Table 9**, five reviews examined the *educational settings*. Meanwhile, for those which presented no explicit indication, the educational setting of the study can be deduced from the intervention type, such as the reviews from Lu et al. (2022) and Cutumisu et al. (2019). In total, six reviews illustrated the *intervention* categories in different manners. However, information on interventions covers some additional aspects, such as the intervention duration, the pedagogical approach, and the tools used in the intervention (Haseski and Ilic, 2019; de Jong and Jeuring, 2020).

DISCUSSION

RQ1 Towards a More Rigorous Review Method and More Exploration of Different Aspects of CT Assessment in Higher Education

The findings on the bibliographical and methodological characteristics of the reviews show an increased focus on exploring CT assessment in higher education. Most reviews are systematic reviews that adopted existing review methodologies or guidelines, were published in journals or, less commonly, in conferences. However, as far as we are concerned, none of the reviews investigated comprehensively existing reviews or reasoned on the choice of the applied methodology. Though it can be argued that CT education, specifically in the context of higher education, is a relatively young field, due to the number of studies and reviews that have been conducted, we contend that an umbrella review is necessary before conducting a systematic review so as to properly locate potential gaps in the field.

Since studies of CT fall under a multidisciplinary scope, the searched databases varied from generic indexing databases such as Scopus to domain-specific databases such as ERIC. With these results, for reviews that aim to examine studies conducted on the topic of CT, it is advisable to include generic databases, as well as domain-specific databases to ensure the maximum exhaustiveness of the search. The retrieved articles, in most of the reviews, were filtered with explicit inclusion and exclusion criteria; however, the quality of the review process was mostly justified by explaining the method used for resolving discrepancies. To have a more rigorous method for reviews, we contend that it is worth discussing the methods to assure the coverage of retrieved data as exhaustively as possible, including the selection of databases, the establishment of inclusion and exclusion criteria, and the resolution of discrepancies.

Topic-wise, only one review investigated CT assessment in higher education, while the rest either studied the topic as complementary aspects of the main focus or with an emphasis on other aspects in higher education. Since most of the reviews aim at discussing CT assessment in higher education as a complementary topic, it should be noted that the insights in this umbrella review may also concern mixed findings, including existing knowledge in K-12 settings as well as in higher education, and other aspects of CT education. Specifically, only some reviews discussed results on assessment constructs, assessment methodologies, and assessment contexts; more effort is necessary to enrich the field from all aspects. However, we consider our findings as a reference for the design, scoping, and discussion of future studies.

RQ2 Assessment Design – Assessment Constructs, Methodologies, and Contexts

For RQ2, we investigated the assessment constructs, assessment methodologies, and assessment contexts of CT assessments explored in the reviews. At a high-level, those three aspects were studied in more than two thirds of the reviews, suggesting that they can be crucial components to consider when designing assessments. For those three aspects, a finer level of investigation provides the reader with more details on the potential factors to consider for assessment design.

First of all, we identified 120 unique constructs, with some being clustered from included studies and the others (Cutumisu et al., 2019; Lu et al., 2022) draw on Brennan and Resnick's (2012) CT framework and the hybrid framework (Brennan and Resnick, 2012; Weintrop et al., 2016; Grover et al., 2017). It appears to be that the hybrid framework covers most of the frequent constructs, which indicates a certain level of consensus on the assessed constructs. However, it should be stressed that these results were based on reviews that also included studies in the K-12 context. Moreover, none of the studies examined if certain constructs appeared more often or are considered more appropriate to be assessed at a certain educational level. Therefore, we argue that there is still room for exploring which constructs should be included in various kinds of assessment within specific contexts.

Secondly, in terms of assessment methodology, we found various assessment methods, assessment tools, assessment instruments, and assessment quality indicators. Methods are grouped and presented differently in four reviews and the reviews promoted a combined use of different assessment methods. Besides that, some reviews investigated extensively the tools and instruments and their characteristics, depicting the delivery of assessments and the medium for assessment in a less abstract and more intuitive way. Most instruments were computer-based with non-automatic scoring, indicating a preference for computer-based assessment over paper-based assessment. Moreover, most tools assess both cognitive and non-cognitive skills, which complies with the idea of a more comprehensive understanding of one's competency level. Nonetheless, none of the reviews systematically examined the tools or instruments regarding the level of skills assessed. Further mapping and categorization of tools and instruments and their intended assessment level of skills is necessary for proper progressive assessment design. Last but not the least, most reviews did not examine the validity and the reliability of the assessment, appealing for more attention on this topic.

Lastly, regarding assessment context, we examined four aspects: academic domain, education level, educational setting, and intervention. We found an increased number of studies integrating CT into various disciplines, with a growing attention to non-CS majors. Studies were mostly conducted in a formal educational setting and assessments were mostly conducted with entry-level or lower-level students within a course or curriculum. However, there is no consensus on which assessment is more proper for different education levels and for students from different backgrounds within the context of higher education. Furthermore, though reviews examined characteristics of the interventions applied for the development of CT skills, there was hardly any focus on the relationship between intervention and assessment design. Future studies on assessment could pay attention to constructive alignment, meaning aligning the assessment with the learning objectives and the interventions used in the education scenarios.

CONCLUSION AND FUTURE WORK

To gain insights on the topic of CT assessment in higher education, this study systematically reviewed 11 reviews from the following perspectives: bibliographical features of the reviews; methodological features of the reviews; constructs/objects being assessed; applied assessment methodologies; assessment contexts which provides information about the assessed participants and the learning activities. The major findings of this study are that (1) there is an increase in the amount of studies on the topic of CT assessment in higher education, while more efforts are still needed for exploring different aspects of the topic to further develop the field; (2) there is little discussion on the selection of the review methodology and the selection of databases; (3) there is a plethora of constructs available to be included in assessments; (4) there exists no systematic investigation of the tools or instruments regarding the level of skills being assessed; (5) there are no guidelines for assessing certain constructs with certain methods in different contexts; (6) constructive alignment theory is hardly applied for aligning the interventions with the learning objectives and assessments.

Regardless of the findings presented above, there are still questions to be explored such that we gain more understanding and knowledge for assessing CT in the context of higher education. In terms of assessment constructs, it can be further examined which are necessary to include in assessments in higher education and in different contexts; towards establishing principles for assessment design, the variation of applied assessment constructs can be studied for diverse assessment contexts; finally, in terms of assessment context, a remaining open question is whether assessment conducted in different contexts should vary and how this variation would affect assessment quality.

REFERENCES

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

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32(3), 347–364. https://doi.org/10.1007/BF00138871

Brennan, K. and Resnick, M. (2012). New frameworks for studying and assessing the development of computational thinking, in *Proceedings of the Annual American Educational Research Association Meeting* (pp. 1–25).

Cutumisu, M., Adams, C. and Lu, C. (2019). A scoping review of empirical research on recent computational thinking assessments. *Journal of Science Education and Technology*, 28(6), 651–676. https://doi.org/10.1007/s10956-019-09799-3

de Araujo, A. L. S. O., Andrade, W. L. and Serey Guerrero, D. D. (2016). A systematic mapping study on assessing computational thinking abilities, in *Proceedings of the 2016 IEEE Frontiers in Education Conference* (pp. 1–9). IEEE. https://doi.org/10.1109/FIE.2016.7757678

de Jong, I. and Jeuring, J. (2020). Computational thinking interventions in higher education: A scoping literature review of interventions used to teach computational thinking, in *Proceedings of the 20th Koli Calling International Conference on Computing Education Research* (pp. 1–10). ACM. https://doi.org/10.1145/3428029.3428055

Denning, P. J. and Tedre, M. (2019). *Computational Thinking*. The MIT Press. https://doi.org/10.7551/mitpress/11740.001.0001

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

Fusar-Poli, P. and Radua, J. (2018). Ten simple rules for conducting umbrella reviews. *Evidence-Based Mental Health*, 21(3), 95–100. https://doi.org/10.1136/ebmental-2018-300014

Grover, S., Basu, S., Bienkowski, M., Eagle, M., Diana, N. and Stamper, J. (2017). A framework for using hypothesis-driven approaches to support data-driven learning analytics in measuring computational thinking in block-based programming environments. *ACM Transactions on Computing Education*, 17(3), 14. https://doi.org/10.1145/3105910

Haseski, H. I. and Ilic, U. (2019). An investigation of the data collection instruments developed to measure computational thinking. *Informatics in Education*, 18(2), 297–319. https://doi.org/10.15388/infedu.2019.14

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

Li, N. and Lefevre, D. (2020). Holographic teaching presence: Participant experiences of interactive synchronous seminars delivered via holographic videoconferencing. *Research in Learning Technology*, 28, 2265. https://doi.org/10.25304/rlt.v28.2265

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

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

McMillan, J. H. (2013). *SAGE Handbook of Research on Classroom Assessment*. SAGE. https://doi.org/10.4135/9781452218649

Poulakis, E. and Politis, P. (2021). Computational thinking assessment: Literature review, in P. Anastasiades and N. Zaranis (eds.), *Research on E-Learning and ICT in Education* (pp. 111–128). https://doi.org/10.1007/978-3-030-64363-8_7

Selby, C. C. and Woollard, J. (2014). Computational thinking: The developing definition, in *Proceedings of the SIGCSE 2014*.

Sullivan, F. R. and Heffernan, J. (2016). Robotic construction kits as computational manipulatives for learning in the STEM disciplines. *Journal of Research on Technology in Education*, 48(2), 105–128. https://doi.org/10.1080/15391523.2016.1146563

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

Taslibeyaz, E., Kursun, E. and Karaman, S. (2020). How to develop computational thinking: A systematic review of empirical studies. *Informatics in Education*, 19(4), 701–719. https://doi.org/10.15388/infedu.2020.30

Taylor, J. (2012). Doing your literature review–Traditional and systematic techniques Jill K Jesson doing your literature review–Traditional and systematic techniques, Lydia Matheson Fiona M Lacey £20.99 192pp 9781848601543 1848601549. *Nursing Research*, 19(4), 45. https://doi.org/10.7748/nr.19.4.45.s7

Tedre, M. and Denning, P. J. (2016). The long quest for computational thinking, in *Proceedings of the 16th Koli Calling International Conference on Computing Education Research* (pp. 120–129). ACM. https://doi.org/10.1145/2999541.2999542

Vinu Varghese, V. V. and Renumol, V. G. (2021). Assessment methods and interventions to develop computational thinking–A literature review, in *Proceedings of the 2021 International Conference on Innovative Trends in Information Technology* (pp. 1–7). https://doi.org/10.1109/ICITIIT51526.2021.9399606

Weintrop, D., Beheshti, E., Horn, M. S., Orton, K., Jona, K., Trouille, L. and Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology*, 25(1), 127–147. https://doi.org/10.1007/s10956-015-9581-5

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

Wing, J. M. (2011). Research notebook: Computational thinking—What and why. *The Link Magazine*, 6, 20–23.
Zhang, X. and Specht, M. (2022). A review of reviews on computational thinking assessment in higher education, in *Proceedings of Sixth APSCE International Conference on Computational Thinking and STEM Education 2022 (CTE-STEM)* (pp. 98–103). https://doi.org/10.34641/CTESTEM.2022.472
Zhu, J. and Liu, W. (2020). A tale of two databases: The use of Web of Science and Scopus in academic papers. *Scientometrics*, 123, 321–335. https://doi.org/10.1007/s11192-020-03387-8