Informatics in Education, 2019 Vol. 18 No. 2, 297–319

DOI: 10.15388/infedu.2019.14

An Investigation of the Data Collection Instruments Developed to Measure Computational Thinking

Halil İbrahim HASESKİ[1], Ulaş İLİC[2]

[1]*Manisa Celal Bayar University, Faculty of Education, Department of Computer Education and Instructional Technologies, Manisa, Turkey*
[2]*Pamukkale University, Faculty of Education, Department of Computer Education and Instructional Technologies, Denizli, Turkey*
e-mail: halil.haseski@cbu.edu.tr, uilic@pau.edu.tr

Received: July 2019

Abstract. The aim of the present study was to investigate the properties of paper-and-pencil data collection instruments developed to measure Computational Thinking (CT) based on several variables. Thus, keywords were identified and used in searches conducted in various databases. The outcomes of the search were analyzed based on the inclusion/exclusion criteria and 64 studies that focused on CT measurement were identified. Content analysis findings were classified under several themes. Based the present study findings, it was determined that the number of tools developed to measure CT demonstrated an increasing trend over time. Furthermore, it was found that the above-mentioned studies included mainly tests. Moreover, it was observed that the processes of ensuring validity and reliability were not clearly specified for more than half of the paper-and-pencil data collection instruments designed to measure CT. Based on the findings, several recommendations were presented for future studies and implementations in the related field.

Keywords: computational thinking, assessment tools, content analysis, assessment of computational thinking.

1. Introduction

Information communication technologies, advanced with the innovations in computer science in the current information age, became an integral part of the everyday life. These technologies, included in every walk of life, while facilitating human life on the one hand, lead to certain transformations by affecting social relations, culture and lifestyles on the other. These changes lead to an increasing complexity in daily problems experienced by individuals and in the need to improve problem-solving skills compatible with the requirements of the current era. Thus, Computational Thinking (CT), which was a term first used by Papert (1980) and took the attention of researchers by the study

of Wing (2006) in the following years, was described as a quality that 21st century individuals should possess in order to systematically analyze the problems they face and to develop and implement the most adequate solutions (Denning, 2009; Guzdial, 2008; ISTE, 2016; Juškevičienė & Dagienė, 2018; Wing *et al.*, 2005).

CT became a popular topic that numerous researchers focused on due to its increasing significance (Selby, 2014). Accordingly, various research were conducted to allow the students to acquire CT skills. Therefore, the literature includes studies conducted to develop CT skills through the use of games, computer programs, robots and software programming (Atmatzidou & Demetriadis, 2016; Berland & Wilensky, 2015; Bers, 2010; Denner *et al.*, 2014; Pellas & Peroutseas, 2016), to identify and develop perception and proficiency in CT (Bower *et al.*, 2017; Ling *et al.*, 2017; Park & Jeon, 2015), and to integrate CT in the curriculum and to develop related curricula (Angeli *et al.*, 2016; Bers *et al.*, 2014; Israel *et al.*, 2015). Measurement of the CT skills was considered as the most significant issue in all studies. This was also denoted as an important issue in the literature (de Araujo *et al.*, 2016; Hadad & Lawless, 2015; Shute *et al.*, 2017). Within the above-mentioned scope, it becomes imperative to measure CT skills effectively to allow future studies on CT skills to achieve their objectives.

1.1. *Measuring CT Skills*

Measurement is based on the numerical expression of observed qualities, and therefore makes these observations valid for everyone, and is an indispensable part of scientific approach (Bandalos, 2018; Lester *et al.*, 2014). Similarly, measuring CT skills became highly significant in order to understand the effectiveness of the studies and applications that anticipate the development of CT skills. Thus, the methods and measurement tools used to obtain accurate measurements became critical (Reynolds *et al.*, 2010). On the other hand, the measurement of CT skills includes a principal topical weakness and although there are several methods used to measure CT in the literature, there is no commonly accepted approach (García-Peñalvo & Mendes, 2017; Gonzalez, 2015; Shute *et al.*, 2017). This leads to a challenge for researchers both in measuring the validity and reliability of CT measurement and in comparing the findings obtained with different methods (Grover & Pea, 2013; Kim *et al.*, 2013; Shute *et al.*, 2017).

In literature, several studies on the measurement of CT skills utilized analytical tools such as Dr. Scratch and Alice to determine the performance of fulfilling a submitted task (Denner *et al.*, 2014; Werner *et al.*, 2012). Similarly, there are studies that investigated the development of CT within the context of robotics (Atmatzidou & Demetriadis, 2016; Jaipal-Jamani & Angeli, 2017; Sullivan & Heffernan, 2016). Furthermore, there are studies in the literature that aimed to determine the development in CT skills through interviews and observation (Cetin, 2016; Israel *et al.*, 2015). Moreover, several paper-and-pencil data collection instruments such as scales, questionnaires, tests and rubric forms were utilized to measure CT skills. Related studies

include the utilization of scales to measure the level of CT and the self-efficacy perception in CT instruction (Korkmaz *et al.*, 2017; Özçınar & Öztürk, 2018), other studies utilized questionnaires to measure attitude, perception and self-confidence towards CT (Hutchins *et al.*, 2017; Ling *et al.*, 2017; Yadav *et al.*, 2014), and achievement tests and rubrics to measure students' knowledge and skills in CT (Berland & Wilensky, 2015; Chang, 2017; Jenkins, 2015).

Although there are several approaches in measuring CT skills, paper-and-pencil data collection instruments such as scales, questionnaires, tests and rubric forms are significant due to their frequent use in literature (Atmatzidou & Demetriadis, 2016; Yadav *et al.*, 2014). However, it was considered that the validity and reliability of the above-mentioned data collection tools could lead to more reliable research outcomes in CT. Although various paper-and-pencil data collection instruments were developed to measure CT skills in literature, the lack of research that examined the qualities of these data collection instruments indicates a significant gap. This lack raises curiosity about the efficiencies and qualifications of the data collection tools developed to measure CT which is a multidimensional skill. Similarly, studies conducted with content analysis to further understand CT emphasized the neglected aspects in the measurement of CT skills (Ilic *et al.*, 2018; Kalelioğlu, 2018; Özyurt & Özyurt, 2015; Shute *et al.*, 2017).

Considering that studies conducted with content analysis on education technologies are instructive for new research (Akbulut &; Cardak, 2012; Hew, Kale, & Kim, 2007; Mikropoulos & Natsis, 2011; Shih *et al.*, 2008), it is possible to suggest that examining the qualities of paper-and-pencil data collection tools developed to measure CT skills using content analysis becomes significant in presenting ideas that facilitate the development of more competent measurement tools and determining more effectively the success of the activities carried out to gain the CT skills. Therefore, the present study focused on the investigation of paper-and-pencil data collection instruments developed to measure CT skills. For this aim, following research questions were specified:

1. What is the distribution of CT measurement tools based on years?
2. What is the distribution of CT measurement tools based on the publication characteristics?
3. What is the distribution of CT measurement tools based on the objectives?
4. What is the distribution of CT measurement tools based on the measurement tool type?
5. What is the distribution of CT measurement tools based on the focused sample characteristics?
6. What is the distribution of CT measurement tools based on validity methods?
7. What is the distribution of CT measurement tools based on reliability methods?
8. What is the distribution of CT measurement tools based on the factors and items characteristics?
9. What is the distribution of CT scales based on factor analysis statistics?

2. Method

2.1. *The Research Model*

In the present study which was designed as a systematic literature review, comprehensive database search was conducted on related databases based on the developed search strategies according to the research questions. Subsequent to the analysis of the studies with inclusion/exclusion criteria, the remaining studies were assessed using a control form developed by the authors. Content analysis method was employed in data analysis. This method was preferred due to its proficiency in comparison, classification and correlation of the collected data (Weber, 1990).

2.2. *Search Strategies*

The article search was conducted in Web of Science, ERIC, Science Direct, Scopus, ProQuest, Google Scholar. Along with the keyword "Computational Thinking", additional keywords such as "measure", "assess", "scale", "test" "validity", "reliability" were used based on the capabilities of the database search engines both in English and Turkish languages. The following search strings were used in English:

"Computational thinking" OR "CT" AND "measure'
"Computational thinking" OR "CT" AND "assess'
"Computational thinking" OR "CT" AND "scale'
"Computational thinking" OR "CT" AND "test'
"Computational thinking" OR "CT" AND "validity'
"Computational thinking" OR "CT" AND "reliability'

Furthermore, studies conducted with content analysis in the literature were also taken into consideration and the studies that were considered adequate for the scope of the present study were also included. In case the full text publications were not accessible, journals and corresponding authors were contacted to retrieve the full texts. The article search was carried out between October 17 and 19, 2018. A total of 117 publications were obtained in the search.

2.3. *The Inclusion/Exclusion Criteria*

Similar to all systematic review studies, inclusion/exclusion criteria were determined based on the present study research questions. The criteria that helped to determine the selection of the studies for the systematic review were as follows:

- A topic directly related to CT content,
- Published as an article, proceeding, thesis or book/book chapter,
- Presence of a paper-and-pencil measurement tool developed for CT,
- An accessible full-text publication.

The first criterion was the choice of topic. Due to its content, CT is considered as a multidisciplinary topic that addresses various fields. Thus, the search results could include publications in several research fields. Given such extensive context, studies that were not directly related to CT such as mathematical computational theories and models were excluded from the analysis. The second criterion focused on the analysis of different types of publications. The third criterion was employed to include studies where CT was the main objective and that developed a paper-and-pencil measurement instrument. Hence, it was ensured that the studies that tackled CT as an additional objective and that did not contribute to CT skills measurement were excluded from the present study. The last criterion was related to the presence of a full-text publication and access. The criterion aimed to eliminate publications such as abstract proceedings, which did not include detailed information on the topic. Furthermore, all non-full-text publications were excluded from the study. Besides, bibliographies of the studies that meet the criteria were examined to reach the maximum number of studies and thus, it was aimed to reach different studies about CT measurement. In line with this method, the study conducted by Özçınar & Öztürk, (2018) was reached as an extra. Finally, 64 out of the 117 accessed studies were included in the study.

2.4. *Data Collection Instruments*

The study validity and the reliability are significant considerations. Validity refers to an impartial observation of the phenomenon (Kirk & Miller, 1986), and reliability refers to the reproducibility of research findings (Merriam & Tisdell, 2015). In this context, in order to ensure the validity and reliability of the present study, a Computational Thinking Assessment Tool Control Form (CTATCF) was developed by the authors to investigate the publications that developed paper-and-pencil data collection instruments. In the development process of CTATCF, the essential characteristics that measurement tools should possess were determined based on the research questions. Furthermore, the general topics used in content analysis in the literature were also taken into consideration (Akbulut & Cardak, 2012; Hinkin, 1995; Shih *et al.*, 2008). In the subsequent phase, three measurement and evaluation experts were consulted for their opinion on the CTATCF. The form, finalized based on the expert opinions, is presented in Table 1.

Table 1

CTATCF criteria and criteria options

Criteria	Options
Indexes	SSCI/SCI/SCI-Exp, ProQuest, Other
Journal Titles	
Subjects of Journals	
Years	2010…2018
Type of Publications	Thesis, Article, Book, Book chapter, Proceedings

Continued on next page

Table 1 - continued from previous page

Criteria	Options
Purpose of Instrument	
Type of Instrument	Test, Survey, Rubric, ...
Type of Items	Open ended, Likert, Multiple choice, …
Sample Type	Student, Teacher, …
Sample Size	0–100, 101–200, 201–300, ...
Validity Method	Pilot study, Expert opinion, ...
Reliability Method	Cronbach Alpha, Inter-rater reliability, ...
Number of Factors	
Number of Items	0–10, 11–20, 21–30, ...
Factors	
Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) statistics	KMO, Bartlett's, RMSEA, GFI, ...

2.5. *Data Analysis*

Spreadsheet software was used during the data analysis in the present study. The CTATCF criteria and criteria options were entered in the spreadsheet columns in the software. The revealed options during the analysis were added to the preliminary options. After the completion of the CTATCF using these criteria and options, descriptive statistics such as frequency and percentage were conducted based on the research questions. An article was counted more than once in the analysis when it contained several options. This information was useful for the researchers in classifying and summarizing the obtained data (Lomax & Hahs-Vaughn, 2012). Two authors, who published several articles on CT, worked independently during the data analysis process. In the analysis, Cohen's Kappa statistic was calculated to determine the inter-rater agreement between the encoders and it was found as κ =.94 for the overall analysis, κ =.90 for the objective criterion, κ =.92 for the criterion of validation methods, κ = .95 for the criterion of reliability methods. Hence, it was possible to suggest that there was a high degree of agreement between the encoders (Landis & Koch, 1977). On the other hand, more in-depth controls were conducted by the authors on several studies that did not explicitly and distinctly defined the characteristics included in the CTATCF. However, in cases where agreement could not be achieved, a measurement and evaluation expert was consulted.

2.6. *Limitations*

It is not possible to examine all studies in the literature in a systemic review (Van der Kleij *et al.*, 2015). Thus, the present study has several limitations. The first limitation is related to the database selection, employed inclusion/exclusion criteria, and the search strategies used in the database search engines. Another limitation is the final date when the search was conducted. Categorically, the CTATCF, developed and used for the analysis of the studies by the authors, could as well be considered as a limitation.

3. Results

3.1. *Distributions Based on Years and Trends*

The distribution of studies on CT measurement tools varied based on the year of publication. The annual distribution of the articles is presented in Fig. 1.

The total number of publications in the last 4 years was 46 (71.9%). The total number of publications in the previous 5 years was only 18 (28.1%). Thus, it is possible to suggest that the publication of studies on measurement tools started in 2010, remained stationary until 2015, yet increased after 2015.

3.2. *Distribution Based on the Publication Type, Journal, Journal Field, Index and Research Field*

Findings indicated that 27 studies on measurement tools were proceedings (42.2%), 23 were articles (35.9%), 11 were theses (17.2%), and 3 were book chapters (4.7%). It could be observed that the articles, which constituted a significant portion of these studies, were published in 20 different journals. It was determined that 15 (65.2%) studies were published in journals indexed in SCI/SSCI/SCI-Exp. This finding could be considered as an indication that the studies were published in widely respected scientific journals. It was observed that the most articles were published in ACM Transactions on Computing Education, Computers & Education, and Computers in Human Behavior, two articles each. It was found that these journals, published in the field of measurement tool development, mainly accepted manuscripts in the fields of education and instructional technologies (65%), computer science and technology (20%) and educational sciences (10%).

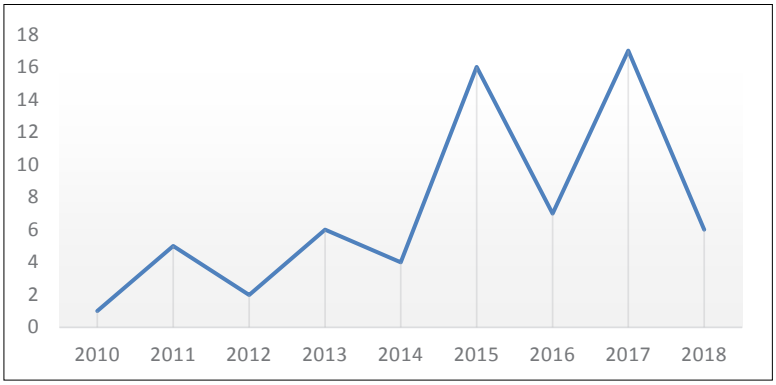

Fig. 1. Distribution of studies based on the year of publication

Table 2
Objectives of measurement instruments

Objective	*f*	%
CT level	30	42.9
CT skills	26	37.1
CT self-efficacy	4	5.7
Attitude towards CT	4	5.7
Self-confidence in CT	1	1.4
Perception of CT	1	1.4
Frequency of selecting the appropriate CT strategy	1	1.4
Effective use of CT in the course plan	1	1.4
Qualification levels of educational digital games for CT	1	1.4
Computational learning (CL) level	1	1.4

3.3. *Distribution Based on the Objectives*

It was determined that the aims of the studies on CT measurement tool development were concentrated in 10 main themes. These objectives are presented in Table 2.

The most significant of the above-mentioned objectives was measuring the CT levels (42.9%). It was followed by the measurement of CT skills (37.1%), the measurement of CT self-efficacy (5.7%), and determination of the attitudes towards CT (5.7%). Given these findings, it was possible to suggest that the studies mostly focused on measuring CT levels and skills (80.0%). Thus, it could be stated that other objectives were neglected.

3.4. *Distribution Based on the Type and Quality of the Measurement Tool*

The analysis demonstrated that the most commonly used type of measurement tool was tests ($f = 36$). It was followed by surveys ($f = 13$), rubrics ($f = 6$) and questionnaires ($f = 5$), respectively. It was found that the number of studies conducted on scale development was quite low ($f = 2$). On the other hand, it was revealed that studied measurement tools mostly included open-ended questions (38.6%). Open-ended questions were followed by Likert-type questions (29.3%) and multiple-choice questions (22.6%). Furthermore, the finding that 4 reviewed studies did not include any discussion on the quality of the measurement tool was noteworthy.

3.5. *Distribution Based on Sample Population and Size*

The majority of the studies conducted to measure CT focused on a target student population (79.7%). Student population was followed by the teacher population (16.4%), one

Table 3
Sample size of the studies

Sample Size	*f*	%
0–100	33	51.6
101–200	15	23.4
201–300	3	4.7
301–400	4	6.3
401–500	1	1.6
501 and higher	4	6.3
Unspecified	4	6.3
Total	**64**	**100.0**

of the major stakeholders in education. The first of the two intended populations included secondary school students (46.8%), high school students (12.6%), primary school students (10.1%) and university students (8.8%). Only one study was conducted with pre-school students, who could be considered to be a significant focus in CT studies. On the other hand, the teacher population included secondary school ($f = 5$) and primary school ($f = 4$) teachers, respectively. In this group, only two studies were conducted with pre-service teachers intended population, and these studies were conducted with pre-school and high-school teachers. It was noteworthy that the number of studies conducted with high school teachers was quite insignificant, despite the frequent selection of high school students. Furthermore, it was also a significant finding that the number of studies conducted with university students and faculty members were insufficient. On the other hand, the fact that pre-school students and teachers were not included as an intended population was another issue that required attention. These findings indicated that the objective that CT was a skill to be adopted by each individual in every aspect of daily life was not fully met. The decrease observed in sample population was observed in the sample size as well.

The study sample sizes are presented in Table 3. The majority of the studies were carried out with a sample size between 0 and 100 subjects (51.6%). This range was followed by the studies that were conducted with sample sizes of 101 and 200 (23.4%). As presented in Table 3, there were studies on measurement tools that required high intended population sizes, however the number of studies that exceeded a sample size of 501 was quite low (6.3%). Furthermore, it was determined that four studies provided no information on the sample size. Thus, it could be stated that the information on the sample was not clearly presented.

3.6. *Distribution Based on the Methods Utilized to Determine Validity*

In the reviewed studies, the methods utilized to determine validity were grouped under 10 themes, as presented in Table 4.

Table 4
Methods to determine validity

Methods	*f*	%
Unspecified	41	56.2
Pilot study	10	13.7
Expert opinion	5	6.8
Construct validity	5	6.8
Content validity	3	4.1
Language validity	3	4.1
Predictive Validity	2	2.7
Face validity	1	1.4
Criterion validity	1	1.4
Concurrent validity	1	1.4
Convergent validity	1	1.4
Total	**73**	**100.0**

As summarized in Table 4, it could be observed that the most common validation method was to conduct a pilot study (13.7%). This method was followed by expert opinion (6.8%), construct validity (6.8%), content validity (4.1%) and language validity (4.1%) methods. On the other hand, it was a noteworthy finding that the majority of studies did not specify which method was used to determine validity (56.2%). Thus, it was possible to suggest that these studies were insufficient in terms of validity. Furthermore, it was determined that the EFA and CFA groups were selected from different groups in two scale development studies analyzed in the present study.

3.7. *Distribution Based on the Methods Utilized to Determine Reliability*

The methods used to estimate reliability in reviewed studies are presented in Table 5. Thus, Cronbach alpha was determined as the most commonly employed method (16.7%). This method was followed by inter-rater reliability (8.3%), test-retest (5.6%) and split-half (4.2%) methods. The analysis results indicated that the majority of the studies did not provide information about the methods used to estimate reliability (59.7%). Thus, it could be argued that more than half of the reviewed studies neglected the reliability factor.

Several methods were employed for item analysis in the reviewed studies in the current study as presented in Table 6. These methods were upper-lower group analysis 27% ($f = 2$), item total correlation ($f = 2$), item effect size ($f = 2$) and item difficulty index ($f = 1$). It was determined that the remaining studies employed none of these methods ($f = 58$). This approach by the majority of the reviewed studies could be interpreted as an evidence that the researchers did not pay sufficient attention to item analysis.

Table 5
Methods to determine reliability

Methods	*f*	%
Unspecified	43	59.7
Cronbach alpha	12	16.7
Inter-rater reliability	6	8.3
Test-retest	4	5.6
Split-half	3	4.2
KR20	1	1.4
Item total correlation	1	1.4
Upper-lower group analysis	1	1.4
McDonald's ω	1	1.4
Total	**72**	**100.0**

Table 6
Methods regarding item analysis

Methods	*f*	%
Upper-lower group analysis of 27%	2	3.1
Item total correlation	2	3.1
Item effect size	2	3.1
Item difficulty index	1	1.5
Unspecified	58	89.2
Total	**65**	**100.0**

3.8. *Distribution Based on the Factors and Number of Items Utilized*

In the present study, the measurement tool factors and item counts were also analyzed. Analysis findings are presented in Table 7.

The majority of the studies included 4 ($f = 11$) or 5 factors ($f = 11$), as presented in Table 7. These structures were followed by 3- ($f = 7$) and 6-factor ($f = 6$) tools, respectively. The fact that a significant number of studies ($f = 15$) did not address the factor count was a significant finding.

When the tolls employed in the studies were analyzed based on item count, it was determined that the majority of the studies included 0 to 10 items (34.4%). This was followed by the studies that included 21 to 30 items (23.4%) and 11 to 20 items (12.5%). In Table 6, it could be observed that the item counts were not specified in 11 reviewed studies. This finding could be interpreted as an evidence that these studies did not provide adequate information on item structures.

Factors employed in reviewed studies are presented in Table 8. As seen in Table 8, the factor frequencies were determined respectively as CT skills ($f = 163$), CT Concepts

Table 7
Number of factors and items

Number of Factors	*f*	%
2	3	4.7
3	7	10.9
4	11	17.2
5	11	17.2
6	6	9.4
7	3	4.7
8	1	1.6
9	2	3.1
10	3	4.7
14	1	1.6
15	1	1.6
Unspecified	15	23.4
Total	**64**	**100.0**

Number of Items	*f*	%
0–10	22	34.4
11–20	8	12.5
21–30	15	23.4
31–40	6	9.4
41–50	2	3.1
Unspecified	11	17.2
Total	**64**	**100.0**

(*f* = 76), CT Patterns (*f* = 28), affective CT achievements (*f* = 9) and cognitive achievements (*f* = 7). The analysis results indicated that among the 47 CT skills, abstraction (*f* = 36), algorithmic thinking (*f* = 12), decomposition (*f* = 12) and sequence of steps (*f* = 10) were the most frequently employed factors. Hence, it was possible to suggest that despite the use of a wide variety of skills, there was an agreement on the significance of certain skills. It was determined that 20 factors were used within the context of CT Concepts. The most commonly used factors were loops (*f* = 15), algorithms (*f* = 12) and conditions (*f* = 7). Therefore, it could be stated that this finding, consistent with the analysis results in CT skills, demonstrated the understanding about the importance of several basic factors. It was determined that 14 factors were used in CT patterns, 8 in affective CT achievements, and 6 in cognitive achievements. The factors in these main themes indicated a homogenous distribution, as presented in Table 8. This finding could be interpreted as a lack of consensus on the factors involved in these themes, in contrast to the heterogeneity experienced with CT skills and CT Concepts.

3.9. *Distribution Based on Factor Analysis Statistics in Scale Development Studies*

The article search, conducted based on the criteria defined in the present study, indicated that there were no scale adaptation studies and only two scales were developed. The fit indices for these studies are presented in Table 9. Initially, it was observed that both studies explicitly reported EFA and CFA statistics. These statistics and the factors included in the studies are listed in Table 9.

It could be stated that the factors used in the study were similar to other measurement tools presented in Table 8 based on the factor count and the factors used in measure-

Table 8
Factors employed in the studies

CT Skills	*f*	CT Skills	*f*	CT Concepts	*f*	CT Patterns	*f*	Cognitive Achievements Towards CT	*f*
Abstraction	36	Connecting	2	Loops	15	Collision	3	CT Term Definition	2
Algorithmic thinking	12	Correctness	2	Algorithm	12	Absorption	3	CT Term Recognition	1
Decomposition	12	Relationship	2	Conditions	7	Generation	3	CT and Other Disciplines Relationship	1
Sequence of Steps	10	Required Task	2	Control Flow	6	Transportation	3	Teaching methods of CT	1
Pattern recognition	6	Logic	1	Variable	6	Diffusion	3	CT Vocabulary	1
Transfer	5	Critical Thinking	1	Debugging	6	Hill climbing	2	Physical Access to CT tools	1
Simulation	5	Reasoning Problems	1	Basic CT Concepts	4	Manipulation	2		
Data Analysis	5	Automating Solutions		Functions	3	Transformation	2		
Parallelization	5	Reuse	1	Data	2	Movement	2	Affective Achievements Towards CT	*f*
Problem Solving	5	Remix	1	Sorting	2	Push	1	Perceived ease of CT integration	2
Evaluation	5	Comparison	1	Searching	2	Pull	1	Attitude Towards CT	1
Collaboration/Cooperation	4	Recall	1	Testing	2	Strategy	1	Motivational Access to CT	1
Identifying Possible Solutions	4	Application	1	Basics of Programming	2	Proximity	1	Perception of Computing	1
Formulating Problems	3	Clarity	1	Boolean Logic	1	Percent chance	1	Perceived usefulness of CT	1
Conditional Logic	3	Efficiency	1	Binary Numbers	1			Self-efficacy	1
Modularity/Modularizing	3	Analogy	1	Cryptology	1			Self-interest	1
Design Solution	3	Mathematical Resolution	1	User input	1			Behavioral Intention	1
Questioning	3	Spatial Reasoning	1	Event handlers	1				
Resources	3	Process	1	Objects	1				
Automation	2	Expressing	1	Thread synchronization	1				
Pattern Generalization	2	Inference	1						
Creativity	2								
Total			**163**	**Total**	**76**	**Total**	**28**	**Total**	**16**

Table 9
Factor analysis statistics

Scales	EFA Statistics	CFA Statistics	Factors
(Korkmaz *et al.*, 2017)	Explained variance= 56.12 KMO= 0.88 Bartlett= 7727.897 (χ^2=7727.897; sd=406; p<0.001)	χ^2/df= 3.232 RMSEA= 0.062 SRMR= 0.044 CFI= 0.95 GFI= 0.91 AGFI= 0.90	1. Creativity 2. Algorithmic Thinking 3. Cooperativity 4. Critical thinking 5. Problem solving
(Özçınar & Öztürk, 2018)	Explained variance= 77.91 KMO= 0.960 Bartlett= 7025.68 (χ^2 =7025.68, p< .01)	χ^2/df= 2.81 RMSEA= 0.08 SRMR= 0.05 NNFI= 0.98 CFI= 0.98 GFI= 0.77	1. Teaching the design of problem 2. Teaching the algorithmic thinking 3. Teaching the evaluation 4. Course planning and teaching methods regarding CT

ments. On the other hand, the researchers stated that EFA and CFA statistics reflected adequate coefficients for the relevant scales (Korkmaz *et al.*, 2017; Özçınar & Öztürk, 2018). Despite the scale studies were infrequent in the literature, it was determined that studies on the development of data collection tools conducted validity and reliability studies meticulously.

4. Conclusions and Further Research

The present study focused on the analysis of 64 studies on the development of paper-and-pencil measurement tools for CT based on their publication years, types, research fields, the aim of the measurement tools developed in these studies, their intended population and population size, the methods adopted to determine validity and reliability, factor and item counts in measurement tools, factors employed in measurement, and EFA and CFA statistics. The present study findings were considered to contribute to future development of quality CT measurement instruments.

Although the concept of CT dates back to the 1980, the present study findings suggested that paper-and-pencil data collection instruments, developed to measure CT skills, emerged only around 2010 in the literature. Thus, it was possible to suggest that CT skill measurement approaches were relatively new. Furthermore, it was determined that the studies on the topic exhibited an increasing trend since 2010. This finding was not consistent with other study results in the literature (Ilic *et al.*, 2018; Şahiner & Kert, 2016). However, it was possible to emphasize that the topic of CT measurement gained significance and an increasing number of researchers started to investigate the topic (Lim, 2015; Selby, 2014; Weintrop, 2016).

The paper-and-pencil data collection instruments that were developed to measure CT were mainly proceedings and articles. These were followed by theses and book chapters, respectively. Based on the finding that CT was scrutinized in different types of

publications, it was possible to argue that CT was a prevalent subject (Denning, 2009; Guzdial, 2008; ISTE, 2016). On the other hand, it was noteworthy that the number of peer-reviewed articles was relatively lower when compared to that of the proceedings, and the number of dissertations that included the criticism of an expert jury was relatively small. The subject fields of the journals, where the CT measurement studies were published, included the fields of education and instruction technologies, computer science and educational sciences. This finding was not consistent with the fact that CT is a skill that should be included in every aspect of life (Ilic *et al.*, 2018; National Research Council, 2010; Wing, 2006).

It was determined that the tools designed to measure CT mainly focused on the determination of CT levels and measurement of CT skills (Basu *et al.*, 2017; Berland & Wilensky, 2015; Chen *et al.*, 2017; Grover *et al.*, 2015; Zhong *et al.*, 2016). There are relatively limited number of studies that focused on measuring CT based on affective variables such as self-efficacy, attitude, self-esteem and perception. Furthermore, it was established that the number of measurement tools designed to determine adequate educational CT strategies, the effective use of CT in syllabi, and the competency levels of educational games in CT development was quite low (Bort & Brylow, 2013; Haseski *et al.*, 2017; Wolz *et al.*, 2011). Although the objective of CT level and skill measurement was prioritized, it was considered that conducting measurements based on different dimensions and contexts could provide a more comprehensive information about CT.

It was observed that the tools developed to measure CT mainly included tests. The reason for the extensive use of tests to measure CT was considered to be due to the ease and low cost associated with test development and application (Lane *et al.*, 2016; Murchan & Shiel, 2017; Reynolds *et al.*, 2009). This finding was inconsistent with the previous research findings which indicated the use of questionnaires as a prominent measurement tool in the field of educational technologies. (Goktas *et al.*, 2012; Hew *et al.*, 2007; Simsek *et al.*, 2009). It was also established that the least used data collection tools included scales, checklists and inventories. Although valid and reliable measurements could be conducted with all utilized data collection instruments, the number of scales structurally validated with statistical analyzes were quite low in the literature, thus, this was considered as a certain problem in CT measurement (Ilic *et al.*, 2018). Furthermore, open-ended questions were frequently used to measure CT. Such an approach stemmed from the fact that open-ended questions were easy to develop and could facilitate data collection from the participants (Fein, 2012; Haladyna & Rodriguez, 2013; Wright, 2008). Open-ended questions were followed by Likert-type and multiple-choice questions, respectively. It was considered that these question types were preferred due to their intermittent measurement properties and provision of simplicity in answering and scoring (Coulacoglou & Saklofske, 2017; Domino & Domino, 2006; Groth-Marnat & Wright, 2016). On the other hand, it was considered that four reviewed studies, which did not indicate the nature of the questions utilized, negatively affected the repeatability of these studies by other researchers.

In the reviewed studies, the intended population mainly included students. This finding was consistent with the findings in studies conducted with content analysis in the

literature (Hsu *et al.*, 2018; Ilic *et al.*, 2018; Shute *et al.*, 2017; Şahiner & Kert, 2016). It was suggested that the reason for insistence on the students as the intended population was due to ease of access and the aim to deliver CT skills at early ages in the formal education process. Besides, it was a remarkable finding that preschool students were the least preferred population in the study samples. This approach was not consistent with the CT approach, which is important for individuals of all ages (Williamson, 2016; Wing *et al.*, 2005). It was established that teachers were the second most preferred intended population, following the students. Similarly, Kalelioğlu (2018) expressed that teachers were selected as the intended population, following the students. On the other hand, it was determined that there were quite few studies that focused on faculty members, high school and preschool teachers as their intended population. Similarly, it was determined that the number of studies that focused on pre-service teachers as their intended population was only a few. It was an interesting finding that high school students were often preferred as the intended population, while high school teachers were hardly included in the studies. In other respects, it was a noteworthy finding that pre-school students and teachers were not selected as the intended population. It could be stated that this finding was not consistent with the idea that CT was a skill that should be acquired by every individual in all aspects of daily life (Cecilia Martinez & Emilia Echeveste, 2015; Lu & Fletcher, 2009). Although, the fact that the selection of students in different age groups was consistent with the idea that CT was significant for individuals of all ages (National Research Council, 2010; Wing, 2008), non-inclusion of several age groups contradicted this fact.

It was determined that studies were conducted mainly with a sample size of up to 100 subjects. It was concluded that the number of studies conducted with larger sample sizes than 200 individuals was relatively small. This finding was consistent with the findings of previous research that investigated CT studies using content analysis in literature (Ilic *et al.*, 2018; Şahiner & Kert, 2016). Four studies included no information on sample size. This approach contradicted with the requisite of specifying sample size in the methodology section in the studies (Henson & Roberts, 2006).

Various methods were used to ensure the validity of the data collection tools developed to measure CT. On the other hand, a major shortcoming was identified as the lack of statements on the means used to ensure validity in more than half of the reviewed studies. Furthermore, several studies that indicated that they consulted expert opinions for validity lacked adequate information on the details of the process of obtaining expert opinion, thus their validity was affected negatively. Considering that the validity of the measurement tool is a prerequisite for accurate measurements (Borsboom *et al.*, 2004; Drost, 2011; Maraun, 2012), it was possible to suggest that half of the developed data collection tools were inadequate for conducting valid measurements.

It was found that several studies reported adequate reliability coefficients for the data collection tools to measure CT (DeVellis, 2012; Huck, 2008; Kline, 2000). On the other hand, a major shortcoming was determined in more than half of the reviewed studies, which entailed failure to specify the method that was used to determine reliability. The reliability of a measurement tool is considered as an absolute prerequisite to conduct accurate measurements (Krippendorff, 2008; Murphy & Davidshofer, 2005; Urbina,

2014). Thus, it was possible to suggest that half of the developed data collection tools were inadequate to conduct reliable measurements.

It was determined that the factor counts in the data collection tools developed to measure CT ranged between 2 and 15. This finding was consistent with the idea that CT should have a multi-component structure (Basawapatna *et al.*, 2011; Román-González *et al.*, 2017; Shute *et al.*, 2017). In other respects, one of the four reviewed studies did not report the measurement factors and did not provide adequate information on the related data collection tools. This finding was considered to negatively affect the validity and reliability of the related research. Furthermore, it was concluded that data collection tools with up to 30 items were preferred frequently. Similarly, it was reported that the scales developed in social sciences included less than 30 items (Hinkin, 1995; Morgado *et al.*, 2018). On the other hand, it was observed that number of items was not reported in several reviewed studies that focused on measurement tool development. Thus, it could be argued that there was not adequate information on data collection tool item structures.

It was observed that the factors in the data collection tools designed to measure CT were mainly based on CT skills (Atmatzidou & Demetriadis, 2016; Bers *et al.*, 2014; Kazimoglu, 2013; Weese, 2017). It was considered that this was due to the reflection of the frequent use of CT in practice. CT Skills was followed by CT Concepts and CT Patterns (Angeli & Jaipal-Jamani, 2018; Basawapatna *et al.*, 2011; Penmetcha, 2012; Worrell *et al.*, 2015). It was considered that the reason for this was based on the fact that programming and computer applications use was predominant in CT measurement, which was based on programming and computer sciences (Israel *et al.*, 2015; Kim *et al.*, 2013; Pellas & Peroutseas, 2016). In other respects, findings revealed that factors on affective achievements in CT and cognitive achievements in CT were the least used factors (Bean *et al.*, 2015; Hutchins *et al.*, 2017; Özçınar & Öztürk, 2018; Webb, 2013). It was considered that this was due to the predominance of application skills in CT measurement, the theoretical knowledge on the concept, and the insignificance of the sensory achievements. Considering the fact that affective features in a situation, concept, or case are important for individuals to take action, it was possible to assert that a relatively low measurement of psychological characteristics in CT could be considered as a problem, since these traits could support the individuals to voluntarily develop CT skills.

The findings indicated that the frequency of the codes under each factor used to measure CT were similar. Thus, it was determined that abstraction, algorithmic thinking, decomposition and sequence of steps were commonly included in CT skills (Atmatzidou & Demetriadis, 2016; Djambong & Freiman, 2016; Weese & Feldhausen, 2017). Furthermore, loops, algorithms and conditional structures were prevalent within the scope of CT Concepts (Bort & Brylow, 2013; Fadjo, 2012; Grover *et al.*, 2014; Román-González *et al.*, 2017). It was considered that the reason of these approaches stemmed from the fact that certain structures and skills were considered more important, although there was no consensus on using the programming structures and skills in measuring CT, which was based on programming (García-Peñalvo, 2018; Kazimoglu *et al.*, 2012). However, CT patterns, cognitive achievements in CT and affective achievements in CT

were the least studied factors (Bean *et al.*, 2015; Leonard *et al.*, 2015; Webb, 2013). This was due to the fact that literature did not exhibit a dominant trend in factor selection. This approach was considered as an important limitation for CT measurement using various dimensions and its comprehension.

Two CT measurement scales were identified in the literature. These scales were "Computational Thinking Scale", developed by Korkmaz *et al.* (2017) and "The Scale for Self-Efficacy Perception Towards Computational Thinking Instruction", developed by Özçınar and Öztürk (2018). Although paper-and-pencil scales were infrequent in the literature, the above-mentioned data collection tools included meticulous validity and reliability studies and met relevant criteria (Thompson, 2004; Snedecor & Cochran, 1989; Pallant, 2001; Byrne, 2016; Hooper *et al.*, 2008). On the other hand, it was considered that limited and valid scale studies on CT measurement were a major gap in the literature. Thus, it could be suggested that the factors that these scales were based on were the most commonly studied factors in the literature and consistent with the present study findings.

Recommendations for future studies:

1. It is essential to develop data collection instruments that effectively meet the validity and reliability criteria and report these qualifications in detail and the studies should clearly report the research methodology utilized in CT studies.
2. Data collection tools that measure CT in various contexts such as educational digital games, CT integration in course plan, and affective traits should be developed.
3. It is important to study various populations to measure CT. Hence, in further studies on CT, valid and reliable data collection tools should be developed to measure CT skills of individuals particularly at kindergarten and university level, employed in different professions, and elderly individuals.
4. It is necessary to conduct studies that would adopt a larger sample size and with higher generalizability in CT measurement.
5. It is essential to conduct studies that focus on the consistencies between the paper-and-pencil measurement tools developed for CT and other methods used to measure CT.
6. Similar studies could be conducted to examine the validity and reliability of CT measurement methods other than the paper-and-pencil applications.

References

Akbulut, Y., & Cardak, C. S. (2012). Adaptive educational hypermedia accommodating learning styles: A content analysis of publications from 2000 to 2011. *Computers & Education, 58*(2), 835–842.

Angeli, C., & Jaipal-Jamani, K. (2018). Preparing pre-service teachers to promote computational thinking in school classrooms. In M. S. Khine (Ed.), *Computational thinking in the stem disciplines* (pp. 127–150). Switzerland: Springer.

Angeli, C., Voogt, J., Fluck, A., Webb, M., Cox, M., Malyn-Smith, J., & Zagami, J. (2016). A K-6 computational thinking curriculum framework: Implications for teacher knowledge. *Educational Technology & Society, 19*(3), 47–58.

Atmatzidou, S., & Demetriadis, S. (2016). Advancing students' computational thinking skills through educational robotics: A study on age and gender relevant differences. *Robotics and Autonomous Systems, 75*, 661–670.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. NY: The Guilford Press.

Basawapatna, A., Koh, K. H., Repenning, A., Webb, D. C., & Marshall, K. S. (2011). *Recognizing computational thinking patterns.* Paper presented at the 42nd ACM technical symposium on Computer science education.

Basu, S., Biswas, G., & Kinnebrew, J. S. (2017). Learner modeling for adaptive scaffolding in a computational thinking-based science learning environment. *User Modeling and User-Adapted Interaction, 27*(1), 5–53.

Bean, N., Weese, J., Feldhausen, R., & Bell, R. S. (2015). *Starting from scratch: Developing a pre-service teacher training program in computational thinking.* Paper presented at the Frontiers in Education Conference (FIE).

Berland, M., & Wilensky, U. (2015). Comparing virtual and physical robotics environments for supporting complex systems and computational thinking. *Journal of Science Education and Technology, 24*(5), 628–647.

Bers, M. U. (2010). The TangibleK robotics program: Applied computational thinking for young children. *Early Childhood Research & Practice, 12*(2), 1–20.

Bers, M. U., Flannery, L., Kazakoff, E. R., & Sullivan, A. (2014). Computational thinking and tinkering: Exploration of an early childhood robotics curriculum. *Computers & Education, 72*, 145–157.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological review, 111*(4), 1061–1071.

Bort, H., & Brylow, D. (2013). *CS4Impact: measuring computational thinking concepts present in CS4HS participant lesson plans.* Paper presented at the 44th ACM technical symposium on Computer science education.

Bower, M., Wood, L. N., Lai, J. W., Howe, C., Lister, R., Mason, R., . . . & Veal, J. (2017). Improving the computational thinking pedagogical capabilities of school teachers. *Australian Journal of Teacher Education, 42*(3), 53–72.

Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (3 ed.). NY: Routledge.

Cecilia Martinez, M., & Emilia Echeveste, M. (2015). Primary and secondary school students' representation about computer sciences and their job. *RED-Revista De Educacion A Distancia*(46).

Cetin, I. (2016). Preservice teachers' introduction to computing: Exploring utilization of scratch. *Journal of Educational Computing Research, 54*(7), 997–1021. doi:http://dx.doi.org/10.1177/0735633116642774

Chang, C. (2017). *Transforming video gameplay experiences into a roadmap to facilitate children's learning of computational thinking concepts.* (Doctoral dissertation), Columbia University, USA.

Chen, G., Shen, J., Barth-Cohen, L., Jiang, S., Huang, X., & Eltoukhy, M. (2017). Assessing elementary students' computational thinking in everyday reasoning and robotics programming. *Computers & Education, 109*, 162–175.

Coulacoglou, C., & Saklofske, D. H. (2017). *Psychometrics and psychological assessment: Principles and applications*. UK: Academic Press.

de Araujo, A. L. S. O., Andrade, W. L., & Guerrero, D. D. S. (2016). *A systematic mapping study on assessing computational thinking abilities.* Paper presented at the 2016 IEEE Frontiers in Education Conference.

Denner, J., Werner, L., Campe, S., & Ortiz, E. (2014). Pair programming: Under what conditions is it advantageous for middle school students? *Journal of Research on Technology in Education, 46*(3), 277–296.

Denning, P. J. (2009). The profession of IT beyond computational thinking. *Communications of the ACM, 52*(6), 28–30.

DeVellis, R. F. (2012). *Scale development: Theory and applications*. LA: Sage.

Djambong, T., & Freiman, V. (2016). *Task-based assessment of students' computational thinking skills developed through visual programming or tangible coding environments.* Paper presented at the 13th International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2016).

Domino, G., & Domino, M. L. (2006). *Psychological testing: An introduction* (2 ed.). UK: Cambridge University Press.

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and Perspectives, 38*(1), 105–123.

Fadjo, C. L. (2012). *Developing computational thinking through grounded embodied cognition.* (Doctoral dissertation), Columbia University, USA.

Fein, M. (2012). *Test development: Fundamentals for certification and evaluation*. USA: ASTD Press.
García-Peñalvo, F. J. (2018). Computational thinking. *IEEE Revista Iberoamericana de Tecnologías del Aprendizaje (IEEE RITA), 13*(1), 17–19.
García-Peñalvo, F. J., & Mendes, A. J. (2017). Exploring the computational thinking effects in pre-university education. *Computers in Human Behavior, 80*, 407–411.
Goktas, Y., Kucuk, S., Aydemir, M., Telli, E., Arpacik, O., Yildirim, G., & Reisoglu, I. (2012). Educational technology research trends in Turkey: A content analysis of the 2000–2009 decade. *Educational Sciences: Theory and Practice, 12*(1), 191–199.
Gonzalez, M. R. (2015). *Computational thinking test: Design guidelines and content validation.* Paper presented at the EDULEARN15 Conference, Barcelona, Spain.
Groth-Marnat, G., & Wright, A. J. (2016). *Handbook of psychological assessment* (6 ed.). USA: Wiley.
Grover, S., & Pea, R. (2013). Computational thinking in K-12: A review of the state of the field. *Educational Researcher, 42*(1), 38–43.
Grover, S., Pea, R., & Cooper, S. (2015). Designing for deeper learning in a blended computer science course for middle school students. *Computer Science Education, 25*(2), 199–237.
Guzdial, M. (2008). Education paving the way for computational thinking. *Communications of the ACM, 51*(8), 25–27.
Hadad, R., & Lawless, K. A. (2015). Assessing computational thinking. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (3 ed., pp. 1568–1578). USA: IGI Global.
Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. NY: Routledge.
Haseski, H. İ., İlic, U., Tuğtekin, U. (2017). Computational thinking in educational digital games: An assessment tool proposal. In H. Ozcinar, G. Wong, & H. T. Ozturk (Eds.), *Teaching computational thinking in primary education* (pp. 256–287). USA: IGI Global.
Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research common errors and some comment on improved practice. *Educational and psychological measurement, 66*(3), 393–416.
Hew, K. F., Kale, U., & Kim, N. (2007). Past research in instructional technology: Results of a content analysis of empirical studies published in three prominent instructional technology journals from the year 2000 through 2004. *Journal of Educational Computing Research, 36*(3), 269–300.
Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management, 21*(5), 967–988.
Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modeling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods, 6*(1), 53–60.
Hsu, T. C., Chang, S. C., & Hung, Y. T. (2018). How to learn and how to teach computational thinking: Suggestions based on a review of the literature. *Computers & Education, 126*, 296–310.
Huck, S. W. (2008). *Reading statistics and research* (5 ed.). Boston: Pearson.
Hutchins, N. M., Zhang, N., & Biswas, G. (2017). *The role gender differences in computational thinking confidence levels plays in stem applications.* Paper presented at the International Conference on Computational Thinking Education 2017.
Ilic, U., Haseski, H. İ., Tugtekin, U. (2018). Publication trends over 10 years of computational thinking research. *Contemporary Educational Technology, 9*(2), 131–153.
Israel, M., Pearson, J. N., Tapia, T., Wherfel, Q. M., & Reese, G. (2015). Supporting all learners in school-wide computational thinking: A cross-case qualitative analysis. *Computers & Education, 82*, 263–279.
ISTE. (2016). CT leadership toolkit.
`http://www.iste.org/docs/ct-documents/ct-leadershipt-toolkit.pdf?sfvrsn=4`
Jaipal-Jamani, K., & Angeli, C. (2017). Effect of robotics on elementary preservice teachers' self-efficacy, science learning, and computational thinking. *Journal of Science Education and Technology, 26*(2), 175–192.
Jenkins, C. (2015). Poem generator: A comparative quantitative evaluation of a microworlds-based learning approach for teaching English. *International Journal of Education and Development using ICT, 11*(2), 153–167.
Juškevičienė, A., & Dagienė, V. (2018). Computational thinking relationship with digital competence. *Informatics in Education*, *17*(2), 265–284.
Kalelioğlu, F. (2018). Characteristics of studies conducted on computational thinking: A content analysis. In M. S. Khine (Ed.), *Computational Thinking in the STEM Disciplines* (pp. 11–29). Switzerland: Springer.
Kazimoglu, C. (2013). *Empirical evidence that proves a serious game is an educationally effective tool for learning computer programming constructs at the computational thinking level* (Doctoral dissertation), UK.

Kazimoglu, C., Kiernan, M., Bacon, L., & MacKinnon, L. (2012). A serious game for developing computational thinking and learning introductory computer programming. *Procedia – Social and Behavioural Sciences, 47*, 1991–1999.

Kim, B., Kim, T., & Kim, J. (2013). Paper-and-pencil programming strategy toward computational thinking for non-majors: Design your solution. *Journal of Educational Computing Research, 49*(4), 437–459.

Kirk, J., & Miller, M. L. (1986). *Reliability and validity in qualitative research*. CA: Sage Publications.

Kline, P. (2000). *The handbook of psychological testing* (2 ed.). London: Routledge.

Korkmaz, Ö., Çakir, R., & Özden, M. Y. (2017). A validity and reliability study of the Computational Thinking Scales (CTS). *Computers in Human Behavior, 72*, 558–569.

Krippendorff, K. (2008). Reliability. In W. Donsbach (Ed.), *The International Encyclopedia of Communication* (pp. 1–6). USA: John Wiley & Sons.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Lane, S., Raymond, M. R., Haladyna, T. M., & Dowing, S. M. (2016). Test development process. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2 ed., pp. 1–37). NY: Routledge.

Leonard, A. E., Dsouza, N., Babu, S. V., Daily, S. B., Jörg, S., Waddell, C., . . . & Boggs, K. (2015). Embodying and programming a constellation of multimodal literacy practices: Computational thinking, creative movement, biology & virtual environment interactions. *Journal of Language and Literacy Education, 11*(2), 64–93.

Lester, P. E., Inman, D., & Bishop, L. K. (2014). *Handbook of tests and measurement in education and the social sciences* (3 ed.). London: Rowman & Littlefield.

Lim, S. (2015). Designing a literature instruction using big data. *Advanced Science and Technology Letters, 97*, 82–87.

Ling, U. L., Saibin, T. C., Labadin, J., & Aziz, N. A. (2017). Preliminary investigation: Teachers' perception on computational thinking concepts. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 9*(2–9), 23–29.

Lomax, R. G., & Hahs-Vaughn, D. L. (2012). *An introduction to statistical concepts* (3 ed.). New York: Taylor and Francis Group.

Lu, J. J., & Fletcher, G. H. (2009). Thinking about computational thinking. *ACM SIGCSE Bulletin, 41*(1), 260–264.

Maraun, M. D. (2012). Validity and measurement. *Measurement: Interdisciplinary Research and Perspectives, 10*(1–2), 80–83.

Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative research: A guide to design and implementation*. NY: John Wiley & Sons.

Mikropoulos, T. A., & Natsis, A. (2011). Educational virtual environments: A ten-year review of empirical research (1999–2009). *Computers & Education, 56*(3), 769–780.

Morgado, F. F., Meireles, J. F., Neves, C. M., Amaral, A. C., & Ferreira, M. E. (2018). Scale development: Ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexão e Crítica, 30*(1), 3–20.

Murchan, D., & Shiel, G. (2017). *Understanding and applying assessment in education*. UK: SAGE.

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications* (6 ed.). NJ: Pearson/Prentice Hall.

National Research Council. (2010). *Report of a workshop on the scope and nature of computational thinking*. Washington: The National Academies Press.

Özçınar, H., & Öztürk, E. (2018). The scale of self-efficacy perception towards teaching computational thinking: A validity and reliability study. *Pamukkale University Journal of Social Sciences Institute, 30*, 173–195.

Özyurt, Ö., & Özyurt, H. (2015). Learning style based individualized adaptive e–learning environments: Content analysis of the articles published from 2005 to 2014. *Computers in Human Behavior*, 52, 349–358.

Pallant, J. (2001). *SPSS survival manual*. PA: Open University Press.

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

Park, S., & Jeon, Y. (2015). Teachers' perception on computational thinking in science practices. *International Journal of Education and Information Technologies, 9*, 180–185.

Pellas, N., & Peroutseas, E. (2016). Gaming in Second Life via Scratch4SL: Engaging high school students in programming courses. *Journal of Educational Computing Research, 54*(1), 108–143.

Penmetcha, M. R. (2012). *Exploring the effectiveness of robotics as a vehicle for computational thinking.* (Doctoral dissertation), Purdue University, USA.

Reynolds, C. R., Livingston, R. B., & Willson, V. L. (2009). *Measurement and assessment in education* (2 ed.). USA: Pearson.

Reynolds, C. R., Livingston, R. B., Willson, V. L., & Willson, V. (2010). *Measurement and assessment in education*. Upper Saddle River: Pearson Education International.

Román-González, M., Pérez-González, J. C., & Jiménez-Fernández, C. (2017). Which cognitive abilities underlie computational thinking? Criterion validity of the computational thinking test. *Computers in Human Behavior, 72*, 678–691.

Selby, C. C. (2014). *How can the teaching of programming be used to enhance computational thinking skills?* (Doctoral dissertation), University of Southampton, Southampton.

Shih, M. L., Feng, J., & Tsai, C. C. (2008). Research and trends in the field of e-learning from 2001 to 2005: A content analysis of cognitive studies in selected journals. *Computers & Education, 51*(2), 955–967.

Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review, 22*, 142–158.

Simsek, A., Ozdamar, N., Uysal, O., Kobak, K., Berk, C., Kılıcer, T., & Cigdem, H. (2009). Current trends in educational technology research in Turkey in the new millennium. *Educational Sciences: Theory & Practice, 9*(2), 941–996.

Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods* (8 ed.). USA: Iowa State University Press.

Sullivan, F. R., & 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.

Şahiner, A., & Kert, S. B. (2016). Examining studies related with the concept of computational thinking between the years of 2006–2015. *European Journal of Science and Technology, 5*(9), 38–43.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington DC: American Psychological Association.

Urbina, S. (2014). *Essentials of psychological testing*. NJ: John Wiley & Sons.

Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of educational research, 85*(4), 475–511.

Webb, H. C. (2013). *Injecting computational thinking into computing activities for middle school girls.* (Doctoral dissertation), Pennsylvania State University, USA.

Weber, R. P. (1990). *Basic content analysis* (2 ed.). CA: Sage.

Weese, J. L. (2017). *Bringing computational thinking to K-12 and higher education* (Doctoral dissertation), Kansas State University, USA.

Weese, J. L., & Feldhausen, R. (2017). *Stem outreach: Assessing computational thinking and problem solving*. Paper presented at the 2017 ASEE Annual Conference & Exposition, Ohio.

Weintrop, D. (2016). *Modality matters: Understanding the effects of programming language representation in high school computer science classrooms.* (Doctoral Dissertation), Northwestern University,

Werner, L., Denner, J., Campe, S., & Kawamoto, D. C. (2012). *The fairy performance assessment: Measuring computational thinking in middle school.* Paper presented at the 43rd ACM Technical Symposium on Computer Science Education (SIGCSE '12), New York.

Williamson, B. (2016). Political computational thinking: Policy networks, digital governance and 'learning to code'. *Critical Policy Studies, 10*(1), 39–58.

Wing, J., Henderson, P., Hazzan, O., & Cortina, T. (2005). Computational thinking. http://www.cs.cmu.edu/afs/cs/usr/wing/www/ct-paper.pdf

Wing, J. M. (2006). Viewpoint: Computational thinking. *Communications of the ACM, 46*(3), 33–35.

Wing, J. M. (2008). Computational thinking and thinking about computing. *Philosophical transactions of the royal society of London A: mathematical, physical and engineering sciences, 366*(1881), 3717–3725.

Wolz, U., Stone, M., Pearson, K., Pulimood, S. M., & Switzer, M. (2011). Computational thinking and expository writing in the middle school. *ACM Transactions on Computing Education (TOCE), 11*(2), 1–22.

Worrell, B., Brand, C., & Repenning, A. (2015). *Collaboration and computational thinking: A classroom structure.* Paper presented at the 2015 IEEE Symposium USA.

Wright, R. J. (2008). *Educational assessment: Test and measurement in the age of accountability*. CA: SAGE Publications.

Yadav, A., Mayfield, C., Zhou, N., Hambrusch, S., & Korb, J. T. (2014). Computational thinking in elementary and secondary teacher education. *ACM Transactions on Computing Education (TOCE), 14*(1), 1–16.

Zhong, B., Wang, Q., Chen, J., & Li, Y. (2016). An exploration of three-dimensional integrated assessment for computational thinking. *Journal of Educational Computing Research, 53*(4), 562–590.

H.İ. Haseski is an assistant professor at the Department of Computer Education and Instructional Technology at Manisa Celal Bayar University, Turkey. He has a Ph.D. in Computer Education and Instructional Technology. His research interests are computational thinking, lifelong learning, e-learning, distance education and social networks.

U. İlic is a research assistant at the Department of Computer Education and Instructional Technology at Pamukkale University, Turkey. He has a Ph.D. in Computer Education and Instructional Technology. His research interests are computational thinking, multimedia learning and the implementation of desirable difficulties in instructional settings.