Education and Information Technologies (2025) 30:2211–2235
https://doi.org/10.1007/s10639-024-12887-6

# Greek translation, cultural adaptation, and psychometric validation of beginners computational thinking test (BCTt)

**Ioannis Vourletsis**[1] · **Panagiotis Politis**[1]

Received: 17 October 2023 / Accepted: 28 June 2024 / Published online: 13 July 2024


## Abstract

Computational thinking (CT) is regarded as a valuable skill set for the students of the 21st century, fostering problem-solving skills applicable to academic disciplines and everyday problems. Assessing CT involves evaluating the development of its concepts, practices, and perspectives. However, establishing comprehensive and validated assessments across different educational levels remains challenging. The Beginners Computational Thinking Test (BCTt) is a validated tool for assessing CT concepts among primary school students, especially during their first grades (ages 5 to 10). This paper describes the translation, cultural adaptation, and psychometric validation of the BCTt for use with Greek students. The translation process involved both forward and backward translation, while the validity assessment included content and construct validity. The psychometric properties of the adapted scale were also evaluated using Item Difficulty Index, Item Discrimination Index, internal consistency, and test-retest reliability. The results indicated that the Greek version of the BCTt can be used as a reliable and valid tool for assessing the CT skills among students in the three lower grades of primary school, with greater suitability for use among students in the two lower grades. Finally, our findings contribute to improving the existing assessment tools tailored to primary school students while guiding future refinement efforts to enhance overall psychometric quality.

**Keywords** Beginners Computational Thinking test (BCTt) · Assessment · Translation · Cultural adaptation · Psychometric properties

✉ Ioannis Vourletsis
vourlets@uth.gr

Panagiotis Politis
ppol@uth.gr

[1] Pedagogical Department of Primary Education, School of Humanities and Social Sciences, University of Thessaly, Argonafton and Filellinon, Volos 38221, Greece

## 1 Introduction

Computational thinking (CT) refers to the mental processes involved in solving problems, designing solutions, and thinking critically using computational concepts and approaches, applicable to various disciplines without necessarily requiring technology (Hazzan et al., 2020; Wing, 2011). CT was introduced by Papert (1980) but it gained significant attention in 2006, when Wing (2006) defined it as a "universally applicable attitude and skill set everyone, not just computer scientists, would be eager to learn and use" (p. 33). CT's potential to bridge STEM disciplines and foster critical thinking through problem-solving suggests its transformative power for education (Tang et al., 2020). Furthermore, equipping students with CT empowers them to not just use technology, but to leverage its potential to solve real-world problems and contribute meaningfully to society (Resnick & Rusk, 2020; Tissenbaum et al., 2019), aligning perfectly with broader educational goals of fostering critical thinking and responsible citizenship.

Despite numerous approaches suggested for integrating CT in school curricula, a recent systematic literature review by Babazadeh and Negrini (2022) noted that "few indications exist on how to assess CT in compulsory school" (p. 1). Notably, effectively assessing CT skills in younger age groups remains a challenge (Poulakis & Politis, 2021). In addition, existing assessment tools often rely on programming environments (Zapata-Cáceres et al., 2020), which may not be suitable for young students unfamiliar with coding concepts. This gap in CT assessment for younger students may hinder our ability to measure their CT skills development and tailor our CT development teaching strategies accordingly.

Our study directly addresses the challenge of assessing CT skills in younger students by introducing a culturally adapted and validated Greek CT assessment tool specifically designed for young primary school students that does not require a programming environment. Our tool draws upon the widely recognized, original *Beginners Computational Thinking Test* (BCTt· Zapata-Cáceres et al., 2020, 2021), chosen for its proven validity and reliability but also its user-friendly, non-programming interface that is suitable for young students. Importantly, by adapting the BCTt for the Greek context,[1] our study addresses the scarcity of similar CT assessment tools available in the Greek language. Last, the content and format of the BCTt aligns with the specific objectives of our study: assessing the psychometric properties of the Greek adaptation of the BCTt and understanding the development of CT skills in young Greek learners.

[1] For more details and access to the Greek version of the BCTt scale, please refer to the designated repository: https://vourletsis.users.uth.gr/.

## 2 Theoretical background

### 2.1 Computational thinking

CT seems to have its origins in the 1940s, when Polya (1945) outlined his four-step problem-solving method applicable in mathematics and other disciplines. However, it was first documented in Papert's (1980) book *Mindstorms*, referring to a mental skill children acquire through programming without providing more details. It is worth noting that, according to Papert, not only the technical but also the social and affective dimensions of learning are very important. Following Wing's statements in 2006, multiple definitions (generic, operational, and educational or curricular) of CT have been proposed (Román-González et al., 2017).

The three-dimensional model proposed by Brennan and Resnick (2012) constitutes an integral part of the educational and curricular definitions of CT and has significantly influenced future discussions. According to this model (Brennan & Resnick, 2012), CT consists of *concepts* (sequences, loops, parallelism, events, conditionals, operators, and data), *practices* (being incremental and iterative, testing and debugging, reusing and remixing, and abstracting and modularizing), and *perspectives* (expressing, connecting, and questioning). As time progressed, the notion of CT evolved, including broader skills and competencies. However, it is generally accepted that CT involves "formulating problems and their solutions in a way that can be effectively executed by an information-processing agent" (Wing, 2011), a human, a machine, or a combination of both, spanning various disciplines (Grover & Pea, 2018). The research conducted by Annamalai et al. (2022) concluded that CT constitutes a valuable skill set that contributes to general problem-solving skills and that the most critical CT dimensions are abstraction, decomposition, debugging and evaluation, algorithms, and generalization. Finally, since CT regards thinking processes, technology is not necessary for its implementation (Hazzan et al., 2020).

### 2.2 Computational thinking assessment

According to Román-González et al. (2019), although a growing number of assessment methods and tools (aligned with the proposed CT definitions and models) have been developed in recent years, little research has been conducted regarding their validity and the ways to integrate them effectively within educational contexts. Other researchers also point out the necessity for validation and large-scale application of these tools (Lu et al., 2022; Tang et al., 2020) or note that most of them focus on the concepts and practices of CT, neglecting its perspectives (Cutumisu et al., 2019). The results of the recent literature review by Poulakis and Politis (2021) further revealed that most existing assessment strategies target upper elementary or middle school students.

CT assessment tools can be categorized as follows: (a) diagnostic tools, assessing subjects' CT aptitudinal level and used in pretest and posttest terms; (b) summative tools, mainly used as posttests; (c) formative-iterative tools, providing learner feedback; (d) data-mining tools, tracking real-time learner progress; (e) skill transfer tools, assessing students' capacity to apply CT in multiple contexts; (f) perceptions-

attitudes scales, assessing students' perceptions about CT and computer science; and (g) vocabulary assessment, assessing "computational thinking language" (Román-González et al., 2019, p. 83). Among the tools included in the abovementioned categories, very few are independent of a programming environment and make suitable use as both pretest and posttest instruments. Additionally, a limitation of many diagnostic tools is that they are not freely available (Guggemos et al., 2023).

A gap exists, as previously mentioned, in CT assessment tools suitable for young students. While initiatives like the *Bebras Challenge* (bebras.org), a widely used contest-based assessment conducted in many countries internationally including Greece, cater to a broad age range including young students, it's important to consider limitations. The Bebras Challenge, categorized as a skill transfer tool, may require considerable reading comprehension from students in order to understand the challenges presented and this could potentially interfere with or mask CT skills, especially at early stages of development (Zapata-Cáceres et al., 2024). Notably, the Computational Thinking Test (CTt; Román-González, 2015; Román-González et al., 2017) stands out as one of these rare tools that fulfill the dual role of pretest and posttest instruments while maintaining independence from a programming environment, emphasizing its significance due to its validity and reliability, albeit intended for students aged 10 and 16.

### 2.3 Original "Beginners Computational Thinking Test"

BCTt (Zapata-Cáceres et al., 2020, 2021) has been developed for younger students, builds upon CTt and shares its capability to function independently of a programming environment. Similar to CTt, BCTt serves as both a pretest and posttest instrument and "can be used in Primary School students, particularly in first grades (5 to 10 years old)" (Zapata-Cáceres et al., 2020, p. 1913). After the first version was designed for primary school students and tested, an improved one was evaluated with students aged 5 to 12 from three schools in Spain. The revised version introduced improvements and additions, such as an additional answer alternative to each question, refined statements of the questions, new collectible elements, and reformulated questions, for the computational concepts' definitions to become more coherent to them. One of the most significant improvements involved introducing symbolism and adding colored shapes to accommodate color-blind students. Furthermore, an administration protocol has been developed, according to which each computational concept has to be explained orally before administering the test. Detailed examples and instructions regarding each question (see Fig. 1) set are included in the protocol.

After completing the validation process, BCTt consists of 25 items, divided into six sets, each covering a distinct computational concept, as outlined in the table presented in Fig. 2, included in the protocol. For each item, only one of the four possible answers is correct.

The statistical analysis of the results obtained from administering the BCTt indicated the presence of a ceiling effect, implying that the test may be less challenging for older students. Specifically, the average difficulty index for the entire sample was notably high (0.81) and medium (0.71) for the first educational stage. Additionally, it was observed that the BCTt's difficulty was relatively balanced when applied

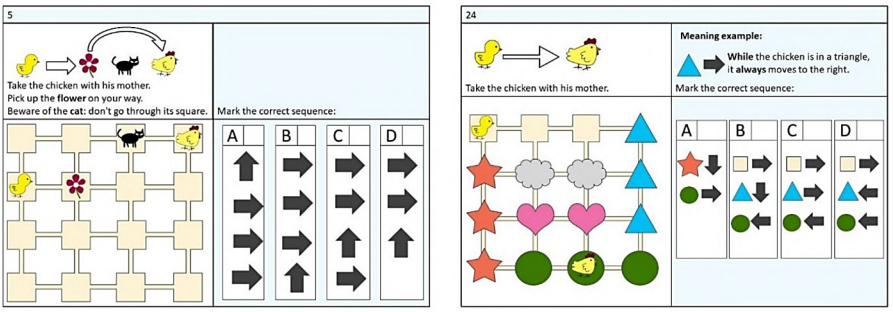


**Fig. 1** Example question in the set of sequences (left) and the set of conditionals (right)

| Item number | Computational concept | | | | | |
|---|---|---|---|---|---|---|
| | SET 1 | SET 2 | SET 3 | SET 4 | SET 5 | SET 6 |
| | Sequence | Loop | | Conditional | | |
| | | Simple | Nested | If-then | If-then-else | While |
| 1 a 6 | x | | | | | |
| 7 a 11 | | x | | | | |
| 12 a 18 | | | x | | | |
| 19 a 20 | | | | x | | |
| 21 a 22 | | | | | x | |
| 23 a 25 | | | | | | x |

**Fig. 2** Computational concept by item

to first grade students ($M=16.59$; $SD=3.104$; $N=70$). Regarding the reliability of the BCTt, it has been excellent for the overall sample ($\alpha=0.824$). However, its reliability decreased with increasing grade levels (first grade: $\alpha=0.833$; second grade: $\alpha=0.793$; fourth grade: $\alpha=0.771$; fifth grade: $\alpha=0.660$; sixth grade: $\alpha=0.657$). This suggests that it is more reliable for the first educational stages. The test-retest reliability analysis, conducted with a 5-week interval, revealed a highly significant positive correlation.

In conclusion, BCTt was sufficient in terms of its design and content. It presents a well-balanced test with progressively increasing difficulty. The first section of the test may prove less challenging for older students, thereby rendering it more suitable for students within the early educational stages (aged 5–10). The reliability of the test was also higher when applied to younger students. Furthermore, it overlooks computational perspectives, focusing more on computational concepts and partially on computational practices. It is independent of any programming environment, allowing it to function as a pretest and posttest tool, facilitating the assessment of multiple dimensions of CT.

## 3 Aim of the study

Given the growing emphasis on CT skills in education, reliable and valid assessment tools are essential for measuring student learning. The primary objective of this study is to assess the psychometric properties of the Greek adaptation of the BCTt, which is used as a valid and reliable assessment tool of the CT skills of primary school students, particularly those aged 5–10. The study focuses on the validity and reliability of the adapted scale, while also examining item analysis to assess the characteristics of individual test items. The specific research questions are as follows:

1. Is the Greek version of the BCTt a valid scale for assessing CT skills for primary school students in Grades 1, 2, 3, and 4?
2. Is the Greek version of the BCTt a reliable scale for assessing CT skills for primary school students in Grades 1, 2, 3, and 4?

The study also aims to contribute to understanding the development of CT skills in young Greek learners while providing educators and researchers with a valid and reliable tool specifically adapted for assessing and promoting CT skills among Greek primary school students.

## 4 Methodology

### 4.1 Participants and data collection

The study employed a two-stage probability sampling technique (see Fig. 3) to obtain a heterogeneous and representative sample of primary school students in Grades 1 to 4 (aged 6–10). During the first stage, we employed a probability proportional to size (PPS) random sampling technique, used when the sampling units vary in size and each unit's inclusion probability needs to be considered (Cheung, 2014). We sorted the 13 Regional Directorates of Primary and Secondary Education in Greece according to their size (number of schools supervised) and randomly selected that of Attica. Right after, we proceeded to the second stage, employing a simple random sampling (SRS) approach, according to which the units are chosen from a population randomly with equal probability (Singh, 2003). After compiling a comprehensive list of schools under supervision, we utilized a random number generator and selected five schools from this list.

Although the necessary sample size to create and validate a scale has been an issue on which there is no consensus, according to Tsang et al. (2017), the respondent-to-item ratio ranges between 5:1 (thus 100 respondents for a 20-item scale) and 30:1. The authors further indicated that "sample sizes of 50 should be considered very poor, 100 as poor, 200 as fair, 300 as good, 500 as very good, and 1000 or more as excellent" (p. S87). As a rule of thumb, it is suggested that a minimum of 10 participants for each scale item are included in the sample or at least 200–300 in total for factor analysis (Boateng et al., 2018). In conclusion, including a larger sample size is always preferable to lower measurement errors and produce results that can be gener-

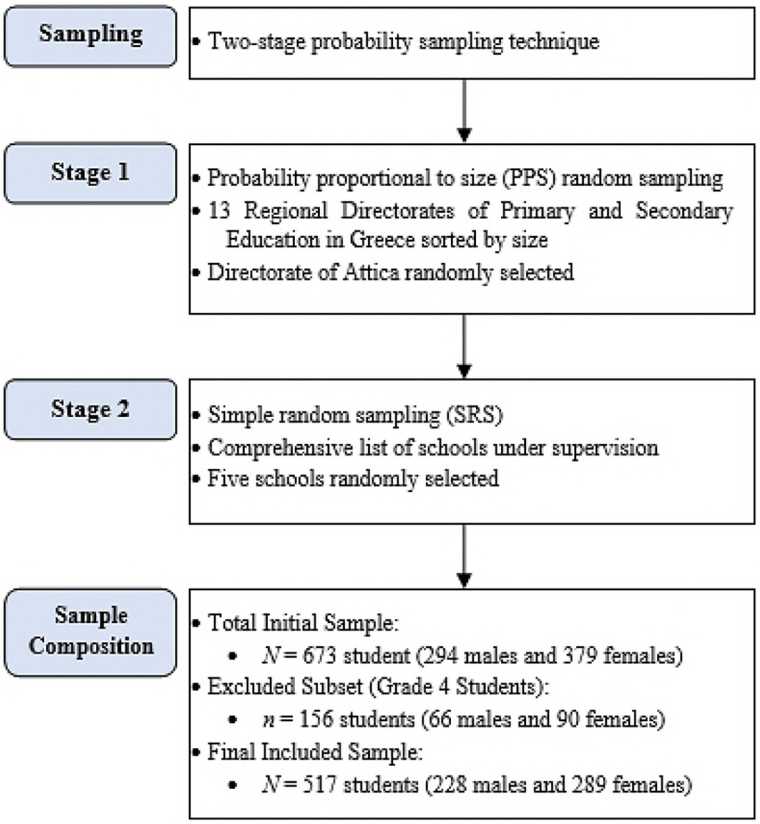


**Fig. 3** Computational concept by item

**Table 1** Distribution of students by grade and gender

| Grade | Male (*n*) | Female (*n*) | Total | Male (%) | Female (%) |
|---|---|---|---|---|---|
| 1 | 63 | 97 | 160 | 39.4 | 60.6 |
| 2 | 77 | 95 | 172 | 44.8 | 55.2 |
| 3 | 88 | 97 | 185 | 47.6 | 52.4 |
| Total | 228 | 289 | 517 | 44.1 | 55.9 |

alized more safely to the actual population (Boateng et al., 2018; Tsang et al., 2017). In the context of factor analysis, researchers commonly utilize the Kaiser–Meyer–Olkin (KMO) Measure of Sampling Adequacy. This statistic indicates the degree to which our data are adequate for factor analysis (Arafat et al., 2016). By employing a combination of PPS and SRS, this study aimed to create a well-balanced and representative sample comprising 673 students (294 males and 379 females). However, while analyzing the scores of the students in each grade, we observed a high prevalence of maximum scores (ceiling effect) regarding the scores of the Grade 4 students. As a result, we excluded those 156 students (66 males and 90 females) from further analysis. Finally, our sample consisted of 517 students, as seen in Table 1.

The first author collected the data from the participating students during May and June 2022, adhering to all ethical considerations. Clear and comprehensive instructions were provided to each class before administering the tests. The instructions included an example of each computational concept included in the test, according to the BCTt protocol. This step was critical to ensuring they were familiar with the assessment format and the concepts evaluated. Following this, students completed the test requiring 35 (Grade 3) to 55 min (Grades 1 and 2). The students' correct answers were coded as 1s and the incorrect answers as 0s using spreadsheet software.

### 4.2 Scale translation and cultural adaptation

For the translation and cultural adaptation of the original BCTt scale into Greek, we adhered to the recommended guidelines (Arafat et al., 2016; Beaton et al., 2000; Boateng et al., 2018; Borsa et al., 2012; Sousa & Rojjanasrirat, 2010; Tsang et al., 2017). This process involved both forward (original language into target language) and backward (target language into original language) translation procedures. However, not consistently following the proposed methodological steps is common between the researchers (Epstein et al., 2015). First, it is essential to point out that the terms "translation" and "adaptation" are distinct. According to Epstein et al. (2015), translation refers to producing a document from a source language into the target language. In contrast, adaptation refers to considering variations between the source and the target culture to preserve equivalence. As a result, we tried to avoid the literal translation of our scale's items and maintain a balance between linguistic, cultural, and scientific information (Borsa et al., 2012) while recognizing the significance of both the target and the source language.

In the forward translation process, two independent bilingual and fully proficient translators made the first translation into their native language (Greek) to more accurately convey the subtleties of the language. To enhance comprehension for the target grade levels, we introduced simplified vocabulary and adjusted the language to ensure the items were more accessible and understandable. Next, we identified, discussed, and resolved the discrepancies between the two translated versions. One more translator, unbiased and bilingual, contributed to the process. According to Epstein et al. (2015), summarizing the scale's translated versions into one is also called reconciliation and seems underestimated.

During reconciliation, adaptations mainly regarded the protagonist's name "chicken," and the phrase "take the chicken with his mother," both present in the majority of the questions within the test. Our discussions focused on these adjustments to align with the audience's age group. Moreover, the test contains minimal text, complementing pictograms and ensuring understanding even without developed reading skills, thus limiting possible adaptations. Besides, most aspects of the original scale were already applicable to Greek language and culture. More specifically, it used culturally neutral elements, including widely recognized visual cues like arrows (for direction) and basic shapes (such as circles, triangles, squares) that transcend cultural barriers. Additionally, marking invalid or inexistent items with an "X" aligns with practices in Greece, demonstrating a shared visual cue. The original character was retained as it does not introduce cultural bias through animal references (chick-

ens are not culturally symbolic or hold negative connotations) for Greek students. Finally, the scale's lack of specific measurement units also facilitated its adaptation to the Greek context.

The synthesized version of the forward translations was then translated into the source language (English) in a process known as backward translation. The latter is considered an "additional quality check" (Borsa et al., 2012, p. 426) and is intended to highlight possible misunderstandings and unclear wordings, thus revealing inconsistencies and conceptual errors (Beaton et al., 2000; Borsa et al., 2012; Tsang et al., 2017). The backward translation was carried out by two independent translators whose native language is the source language of the scale. In addition, to avoid bias, the back-translators were unaware or informed of the scale's concepts. Borsa et al. (2012) state that back-translation requires conceptual equivalence rather than keeping items identical to the original ones. In other words, the two back-translated versions should reflect the content of the original scale's items. During the backward translation process, no further adaptations were introduced, as the focus was on ensuring conceptual equivalence rather than making additional linguistic adjustments.

Next, we produced a prefinal version of the translated scale by reviewing all the translations and determining whether the translated and original versions were semantically and conceptually equivalent. During this phase, no further adaptations or modifications were deemed necessary, as the language adjustments made earlier sufficed to ensure comprehension across different age groups. Finally, we tested the prefinal version of the scale with 43 participants in Grades 1 and 3. According to Beaton et al. (2000), it should be between 30 and 40 participants; other researchers propose participant numbers ranging from 10 to 40 (Sousa & Rojjanasrirat, 2010) or 30 to 50 (Tsang et al., 2017). When applying the prefinal version, the respondents completed the scale and were asked questions to evaluate the appropriateness of the scale's items regarding their meaning and difficulty (Borsa et al., 2012). No modifications were deemed necessary, resulting in the production of the final version of the scale.

### 4.3 Validity

We used the Content Validity Index (CVI) and Confirmatory Factor Analysis (CFA) to evaluate the validity of the translated and culturally adapted BCTt.

#### 4.3.1 Content validity index

To ensure that a scale's items correctly reflect the construct of interest, it is crucial to evaluate its validity. According to Polit and Beck (2006), although various definitions have been proposed, it is generally accepted that content validity refers to "the degree to which a sample of items, taken together, constitute an adequate operational definition of a construct" (p. 490). In other words, content validity is the degree to which the items of a scale are relevant and representative of the intended construct (Yusoff, 2019). The most frequently cited measure of content validity is the CVI, credited to Martuza (1977). To better measure the content validity of our translated scale, we measured the content validity of individual scale items (I-CVI) and the overall scale's

content validity (S-CVI). To calculate the I-CVI, we asked six experts working within the Greek education system to rate the degree to which each item is relevant to the underlying construct using a four-point scale (1=*not relevant*, 2=*somewhat relevant*, 3=*quite relevant*, and 4=*highly relevant*; Lynn, 1986). The panel comprised three university professors specializing in Information and Communication Technologies (ICT) in Education and three Computer Science educators. All Greek native speakers, these experts possess high levels of expertise in Computational Thinking (CT) education. Then, I-CVI was calculated as the proportion of content experts that gave each item a relevant rating of 3 or 4. The S-CVI was calculated as the proportion of the scale's items that achieved a rating of 3 or 4 by the content experts. The average of the I-CVI scores for all items constitutes S-CVI/Ave, but when all the content experts give a rating of 3 or 4, we refer to it as S-CVI/UA (universal agreement; Beck & Gable, 2001; Polit & Beck, 2006; Yusoff, 2019). As recommended, we have reported both S-CVI/Ave and S-CVI/UA (Polit & Beck, 2006). Finally, according to Lynn (1986), the minimum recommended I-CVI values for a scale of excellent content validity are 0.78 for 6 to 10 experts and 0.90 for S-CVI/Ave. The accepted value of S-CVI/UA is equal to or higher than 0.80 (Shi et al., 2012).

#### 4.3.2 Construct validity

To assess the degree to which the adapted scale accurately measures the intended construct (construct validity), we employed factor analysis. Our approach involved employing the CFA, which is a theory-driven method based on an existing hypothesis or "theory," to test if the collected data fit the theoretical model (Hurley et al., 1997). As a result, we used CFA to validate the preexisting CT theoretical framework and factor structure. Furthermore, our data were binary (0 for false and 1 for correct answers). In such cases, it is suggested to use the Weighted Least Squares Means and Variance Adjusted (WLSMV) estimator (Passos et al., 2023). Consequently, we used the WLSMV estimator in R software (R CORE Team, 2019) with the lavaan package (Rosseel, 2012) to conduct our CFA and assess the fit of our hypothesized measurement model. Prior to CFA, KMO test's value was calculated using the psych package in R (Kaiser, 1974) to assess variable sampling adequacy, and Bartlett's test of sphericity confirmed correlations between the variables (Bartlett, 1951). The value of KMO should be higher than 0.60 (Kaiser, 1974) or 0.50 (Hair et al., 2006), and Bartlett's test of sphericity should be statistically significant.

Among the model fit indices we used and reported are the chi-square ($\chi^2$) statistic, root mean square error of approximation (RMSEA), root mean square residual and standardized root mean square residual (SRMR), comparative fit index (CFI), Tucker–Lewis Index (TLI), and normed-fit index (NFI). We also reported the ratio of the $\chi^2$ statistic to its degrees of freedom (df) to provide insight into the model's fitness while considering its complexity. Regarding the cut-off values of the model fit indices, the CFI should be higher than 0.95 (Hu & Bentler, 1999) or 0.90 (Byrne, 1994). The value of TLI should be no less than 0.90 (Bentler & Bonett, 1980) or higher than 0.95 (Hu & Bentler, 1999), and the value of NFI should be higher than 0.90 (Byrne, 1994) or 0.95 (Schumacker & Lomax, 2004). The value of RMSEA should be between 0.05 and 0.08, but values between 0.08 and 0.1 can be accepted

(Fabrigar et al., 1999), and the value of $\chi^2/df$ should be lower than 3 (Kline, 2011) or 5 (Wheaton et al., 1977). Finally, the value of SRMR should be lower than 0.08 (Hu & Bentler, 1999), but SRMR is sensitive to sample sizes and is not recommended for use with binary data (Yu, 2002). It should be noted that although reporting multiple model fit indices is generally recommended, some researchers recommend reporting the TLI, CFI, and RMSEA for one-time analyses and other indices only after modifying the model (Schreiber et al., 2006).

### 4.4 Item analysis

We employed Classical Test Theory (CTT; Hambleton & Jones, 1993) methods to analyze the characteristics of individual items in the Greek BCTt. This analysis focused on item difficulty and discrimination power, using appropriate statistical procedures.

#### 4.4.1 Item difficulty index

After establishing the scale's content validity through the CVI assessment, we continued the psychometric evaluation of the translated scale by examining the item difficulty level. The IDI is used in CTT to describe the difficulty of a single item on a scale. The IDI is defined as the percentage of the group who answer an item correctly (Barnard, 1999) or the relative frequency with which the test takers respond correctly (Thorndike et al., 1991). Accordingly, we calculated the IDI by dividing the total number of individuals who answered correctly by the total number of individuals who answered. The result varies from 0 to 1 and is not often converted into a percentage. Higher IDI values indicate easier questions. To determine if a question is easy, researchers have suggested thresholds, which can be regarded as arbitrary. According to Azevedo et al. (2019), the IDI takes values between 0.15 and 0.85; outside of this range, questions must be reevaluated. Items with IDI values between 0.1 and 0.3 are generally considered difficult, while those with IDI values between 0.7 and 0.9 are generally considered easy (El-Hamamsy et al., 2022).

Furthermore, challenging tests containing items with IDI values below 0.25 tend to be positively skewed, whereas very easy tests with items having IDI values greater than 0.80 tend to be negatively skewed (Nitko & Brookhart, 2014). However, we can place easy items at the start of a test as "warm-up" questions (Hingorjo & Jaleel, 2012, p. 143). Finally, it is generally advised that a good test should contain items with a range of IDI values, and the IDI should ideally be around 0.5 (Azevedo et al., 2019).

#### 4.4.2 Item discrimination index

Our next critical step in the psychometric evaluation of the translated scale was to explore the Item Discrimination Index for each item and the overall scale across grades. The Item Discrimination Index, also called point-biserial correlation, is a method used in CTT to assess the effectiveness of certain test items in differentiating between students of high and low ability (Azevedo et al., 2019; Hingorjo & Jaleel,

2012). More specifically, the Item Discrimination Index uses the differences between the high-ability students who answered the item correctly and the low-ability students who answered correctly and took values between – 1 and 1 since it is a correlation coefficient (Mitra et al., 2009). The value 1 denotes a perfect correlation between the score obtained in a certain question and the total score. As a result, the higher the score on this question, the higher the total test score will be.

Conversely, -1 denotes an inverse perfect correlation, meaning that the higher the score in this question, the lower the total test score. It is generally assumed that individuals of high ability would choose the correct answer to each scale item more often than individuals of low ability. In this case, the item is considered to have a positive Discrimination Index. However, when low-performing individuals answer correctly more often than high-performing individuals, that item has a negative Discrimination Index (Hingorjo & Jaleel, 2012). The threshold of 0.2 is generally applied to retain a scale's items (El-Hamamsy et al., 2022), which is consistent with Ebel and Frisbie's (1991, p. 232) recommendations, as can be seen in Table 2.

### 4.5 Reliability

We assessed the adapted BCTt's reliability using both internal consistency and test-retest reliability.

#### 4.5.1 Internal consistency

In our psychometric evaluation of the translated scale, we also assessed its internal consistency. According to Tsang et al. (2017), internal consistency refers to "the extent to which the questionnaire items are inter-correlated or whether they are consistent in the measurement of the same construct" (Tsang et al., 2017, p. S85). In other words, internal consistency refers to the degree to which a scale's items measure the same underlying concept, skill, or construct (Azevedo et al., 2019), ensuring the accuracy and stability of the scores obtained from a scale. Cronbach's alpha (Cronbach, 1951), also known as coefficient alpha, is a widely used measure of the internal consistency of a scale. A significant value of α is necessary for internal consistency, but it is not guaranteed; lengthy, multidimensional scales will also have high values of α (Streiner, 2003).

Another widely used measure of a scale's internal consistency, mainly when it contains items with only right and wrong answers (dichotomous items) of varying difficulty, is KR20 (Kuder & Richardson, 1937). KR20 is calculated by dividing the average squared correlation between pairs of test items by the average squared correlation of each item with the total test score. It also ranges from 0 to 1, and the scale's internal consistency will be closer to 1, while 0.80 indicates a reasonable consistency

**Table 2** Item discrimination index and its interpretation

| Index of discrimination | Item evaluation |
|---|---|
| 0.40 and above | Very good item |
| 0.30–0.39 | Reasonably good item |
| 0.20–0.29 | Marginal item |
| Below 0.19 | Poor item |

(Azevedo et al., 2019). In our case, for our scale containing binary answers, we calculated the KR20 coefficient for the entire translated scale and its subscales for each student's grade and across all grades. We also need to point out that for our data, which contained only binary data, the value of the KR20 coefficient was identical to Cronbach's alpha coefficient. Finally, we calculated and reported the mean inter-item correlations for the scale and its subscales since Cronbach's α values can often be low with short scales (Streiner, 2003). According to Briggs and Cheek (1986), the inter-item correlations are not influenced by scale length, with the recommended range falling between 0.2 and 0.4.

#### 4.5.2 Test-retest reliability

Test-retest reliability is a complementary approach to evaluating a scale's reliability. It is also known as the coefficient of stability, and we used it to assess the degree to which the participant's performance is repeatable, i.e., the degree to which their answers remain consistent throughout several administrations of a scale (Arafat et al., 2016; Boateng et al., 2018; Tsang et al., 2017). To assess test-test reliability, we administer the same scale twice or more to the same individuals and then calculate Pearson's product-moment correlation coefficient (Pearson's *r*) or the intraclass correlation coefficient (ICC; Tsang et al., 2017). A more significant coefficient indicates more robust test-retest reliability. Broglio et al. (2007) argues that while Pearson's *r* is a bivariate measure of the relationship between two independent variables, ICC is a univariate measure of the score consistency across two different time points. As a result, ICC is often used instead of Pearson's r to assess the consistency of measurements made on the same subject across time or by several raters. ICC values between 0.61 and 0.80 are generally considered moderate, while values between 0.81 and 1.00 are substantial (Vaz et al., 2013).

In our study, we computed the ICC (1, 1), also known as one-way random effects, absolute agreement, single rater or measurement (Koo & Li, 2016), to evaluate the test-retest reliability of our scale between two-time points for each grade as well as for all grades collectively. ICC (1, 1) is generally used to measure the consistency or agreement between measurements taken on the same subjects at different time points or under different conditions. We also reported the confidence intervals of the ICC values. Finally, although the ideal testing interval may vary, 2 weeks is the most recommended interval between the test and retest (Streiner et al., 2015).

## 5 Results

### 5.1 Score analysis

First, we produced descriptive data regarding the students' scores on the adapted scale version. Our results (see Table 3) showed that across all grades ($N=673$), the mean total score is 15.44, with a median of 16.00, which indicates a relatively balanced distribution of scores. Additionally, the value of the standard deviation ($SD=6.20$) indicates that the data are dispersed moderately.

**Table 3** Descriptive statistics of total scores by grade

| Grades | *N* | Mean | Median | Std. Deviation | Minimum | Maximum |
|---|---|---|---|---|---|---|
| 1 | 160 | 12.44 | 13.00 | 6.22 | 0.00 | 24.00 |
| 2 | 172 | 13.49 | 14.00 | 5.96 | 0.00 | 24.00 |
| 3 | 185 | 16.35 | 16.00 | 5.04 | 7.00 | 25.00 |
| 4 | 156 | 19.58 | 21.00 | 5.07 | 6.00 | 25.00 |
| All grades | 673 | 15.44 | 16.00 | 6.20 | 0.00 | 25.00 |

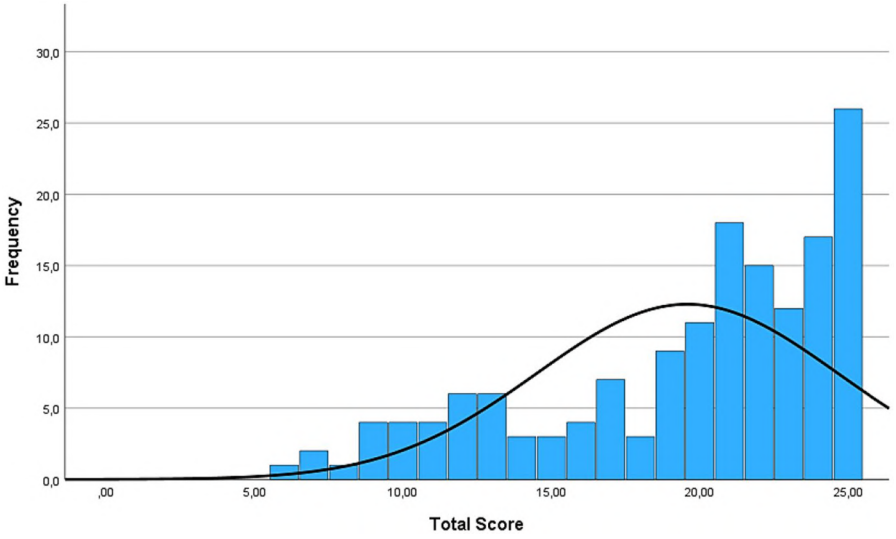


**Fig. 4** Frequency distribution and normal curve fit for Grade 4 data

Looking at individual grades, Grade 4 achieved the highest mean total score (19.58), while 26 participants (approximately 17%) achieved the highest score, 25 points (see Fig. 4). This percentage of high-achieving students led us to observe a ceiling effect, exclude Grade 4 data from further analysis, and subsequently not include them in the data tested for model fit. Our decision was based on the traditionally adopted benchmark of 15% or more of respondents who achieve the lowest or highest possible score, thus signaling a floor or ceiling effect (Terwee et al., 2007).

The mean scores of Grades 1 to 4 indicate an increasing performance across grades, while the value of the standard deviations indicates similar levels of score dispersion within each grade. The standard deviation for Grade 1 is slightly higher, reflecting a broader range of performance levels. In summary, the descriptive statistics reveal variations in total scores across grades, with Grade 4 showing the highest scores and Grade 1 displaying slightly lower scores. The statistics highlight the range and distribution of scores, providing insights into the overall performance of students in CT skills across different educational stages.

## 5.2 Validity

### 5.2.1 Content validity index

As mentioned earlier, a panel of six experts independently rated the degree to which each item was relevant to the underlying construct using a four-point scale. All items received an I-CVI value above 0.78, as Lynn (1986) recommended. The results also indicated unanimous agreement among experts regarding the overall relevance of the scale items since the average proportion of items judged as relevant across the experts (S-CVI/Ave) was 0.99 and the S-CVI/UA was 0.92, suggesting significant consensus regarding the representativeness of the items according to the recommended values (Lynn, 1986; Shi et al., 2012).

### 5.2.2 Construct validity

As described in Sect. 4.3.2, we tested whether our theoretical model fit the data we collected from Grades 1, 2, and 3. Our results of the CFA with six factors based on 25 dichotomous items are summarized in Table 4. We applied the WLSMV estimation method and found that for Grade 1, our model has a good fit according to all indices except for SRMR, whose value was higher than the recommended threshold of 0.09. The $\chi^2/df$ ratio was below 3, which indicates a good fit, and the CFI, TLI, and NFI values are also acceptable. The values of KMO and Bartlett's test of sphericity also indicate that the sampling is adequate and that there is a substantial correlation in the data. We found a similar fit for the model with the data we collected from Grade 2 students, as the $\chi^2/df$ ratio remained below 3, the other indices were within the acceptable range, and the KMO and Bartlett's test of sphericity were as well. However, the value of SRMR exceeded the recommended threshold. The fit indices for Grade 3 indicated a poor fit since only the $\chi^2/df$ ratio was lower than 5, indicating an acceptable fit. The values of all the other indices were not acceptable, and the KMO value showed mediocre sampling.

Finally, we found a good fit when we tested our model with students from all grades. We found an $\chi^2/df$ ratio below 5, which is acceptable; a CFI between 0.90 and 0.95, which is good; good TLI and NFI values as well; and a moderate RMSEA. However, the SRMR value was not good. The sampling was adequate, and there was a substantial correlation in the data, as suggested by the KMO and Bartlett's test of sphericity, respectively. In conclusion, when we tested our model with the data for all grades, we found that Grades 1 and 2 individually exhibit a good fit across all indices, while Grade 3 shows a comparatively less favorable model fit.

**Table 4** Model fit indices comparison and recommended cut-off values for different grades

| Fit index | KMO | Bartlett's test of sphericity | $\chi^2$/df | CFI | TLI | NFI | RMSEA | SRMR |
|---|---|---|---|---|---|---|---|---|
| Grade 1 | 0.84 | $\chi^2$ (300) = 1399, $p < .001$ | 1.870 | 0.954 | 0.949 | 0.907 | 0.074 | 0.144 |
| Grade 2 | 0.82 | $\chi^2$ (300) = 1408, $p < .001$ | 1.994 | 0.944 | 0.937 | 0.895 | 0.076 | 0.146 |
| Grade 3 | 0.60 | $\chi^2$ (300) = 1580, $p < .001$ | 3.463 | 0.787 | 0.749 | 0.728 | 0.121 | 0.196 |
| All grades | 0.82 | $\chi^2$ (300) = 4414, $p < .001$ | 4.914 | 0.934 | 0.920 | 0.918 | 0.087 | 0.128 |

### 5.3 Item analysis

#### 5.3.1 Item difficulty index

We calculated the IDI to describe the difficulty level of each item on our translated scale across grades. Our results (see Fig. 5) showed that in Grade 1, the mean IDI is 0.50, suggesting a relatively balanced distribution of item difficulty levels. However, items with higher values, such as Item 1 (0.90) and Item 6 (0.76), seem easier. On the other hand, Item 21 (0.26) and Item 25 (0.28) demonstrate more significant difficulty levels. In Grade 2, the IDI values show that the scale contains items of varying difficulty. Items like Item 1 (0.95) and Item 3 (0.76) are relatively easier, while Item 25 (0.30) and Item 24 (0.37) exhibit higher levels of difficulty. The mean IDI for Grade 2 is 0.54, indicating a slightly lower overall difficulty compared to Grade 1. However, the mean IDI remains very close to the value of 0.50, as suggested by many researchers (Azevedo et al., 2019). The IDI values for the students in Grade 3 also demonstrated varying difficulty. Some items like Item 1 (0.98) and Item 2 (0.84) appear to be relatively easier, but Item 24 (0.37) and Item 25 (0.38) present higher levels of difficulty. The mean IDI for Grade 3 is 0.63, suggesting an increase in mean IDI compared to Grades 1 and 2, thus an easier test.

The mean IDI for all grades combined is 0.51, indicating a balanced overall distribution of item difficulty levels across the entire scale. The trendline fitted to the difficulty index data points shows that the scale includes items with increasing difficulty. Furthermore, the value of $R^2 = 0.726$ of the trendline suggests that the trendline can explain a relatively high proportion (about 73%) of the variability in the IDI values.

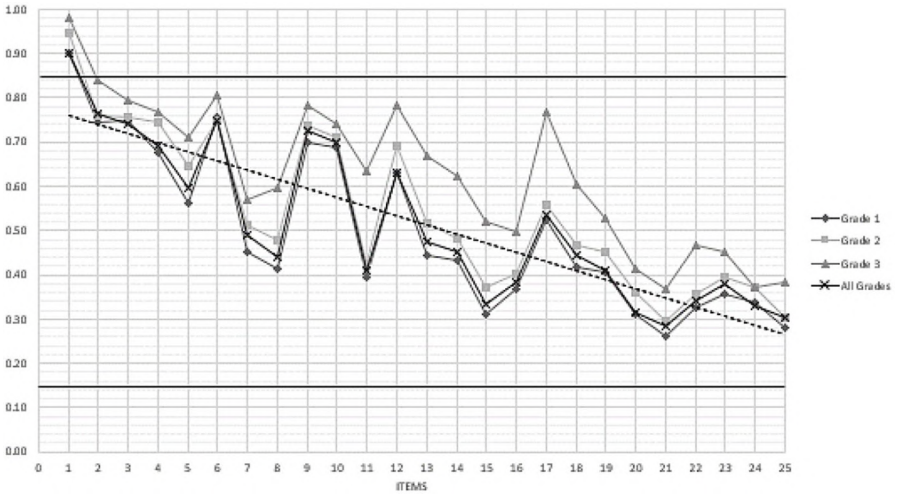


**Fig. 5** Analysis of the item difficulty index with thresholds: variations across grades and overall

### 5.3.2 Item discrimination index

The values of the Item Discrimination Index values of our translated scale showed that it consists of items with varying levels of discriminatory power (see Fig. 6). More specifically, in Grade 1, we found items such as Item 12 (0.62) and Item 22 (0.62) with strong discriminatory power, which can effectively distinguish between students of varying abilities. However, all items fall into the "very good" range, and overall, the mean discrimination index of 0.53 suggests that the assessment is reasonably effective in discriminating between Grade 1 students. In Grade 2, a similar pattern emerged as items showed diverse IDI. Items 2 (0.64) and 18 (0.64) have strong discriminatory power, and 22 out of 25 items fall into the "very good" range. The remaining three items, Item 1 (0.37), Item 9 (0.39), and Item 3 (0.39), fall into the "reasonably good" range. The mean IDI of 0.51 demonstrates a reasonable ability of the assessment to discriminate among Grade 2 students. In Grade 3, the mean IDI was lower (0.41), suggesting that the items can discriminate effectively among Grade 3 students but are less effective than in other grades. Item 1 (0.17) showed poor discriminatory power, but it can be used as a warm-up question. Item 10 (0.25) and Item 11 (0.26) fall into the "marginal" category. While eight items fall into the "reasonably good" category, the remaining 14 items fall into the "very good" category.

In conclusion, the results of the IDI showed that the scale's items can adequately discriminate across students of all grades (0.53) but have greater discriminatory power for Grades 1 and 2 compared to Grade 3. However, we decided not to remove any items from the translated scale, as the individual IDI values exceeded the threshold of 0.20 (El-Hamamsy et al., 2022). Item 1, despite its low IDI value, especially for Grade 3, can be retained and used as an introductory question, given that it is an easy one (Hingorjo & Jaleel, 2012).

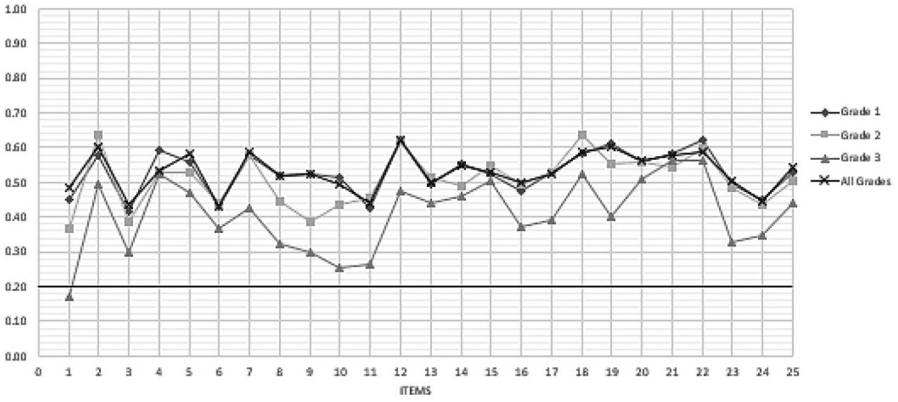


**Fig. 6** Analysis of the item discrimination index with threshold: variations across grades and overall

## 5.4 Reliability

### 5.4.1 Internal consistency

To assess the internal consistency of our translated scale, we calculated the value of the KR20 coefficient and the mean inter-item correlations for the subscales across different grades (see Table 5). The *Sequence* subscale demonstrated adequate internal consistency across all grade levels, and its mean inter-item correlations reflected reasonable associations among items. The *Simple Loop* and *Nested Loop* subscales also showed adequate reliability across grades, and their mean inter-item correlations were relatively higher, signifying stronger linkages among the items. Regarding the *Conditional* subscales, we found low or adequate values of the KR20 coefficient. However, the *Conditional* subscales contained few items, and it is essential to examine the values of the mean inter-item correlations, which indicated satisfactory levels of item relationships and, thus, internal consistency.

The entire scale, comprising 25 items, consistently exhibited high reliability coefficients across all grade levels: Grade 1 (KR20 = 0.89), Grade 2 (KR20 = 0.88), Grade 3 (KR20 = 0.80), and a combination of all grades (KR20 = 0.88). When examined along with the corresponding mean inter-item correlations, these values indicate significant internal consistency of the assessment, encompassing diverse items that evaluate various constructs. The KR20 coefficients and the mean inter-item correlations showed no improvement upon removing any items. Consequently, we retained all the 25 items on the translated scale.

### 5.4.2 Test-retest reliability

We examined the test-retest reliability of the translated scale in a sample of 64 students representing all three grades. We administered the scale to the students in the

**Table 5** KR20 reliability coefficients and Mean Inter-item correlations for the Scale and its Subscales Across grades

| Measure | Grade 1 | | Grade 2 | | Grade 3 | | All Grades | |
|---|---|---|---|---|---|---|---|---|
| | KR20 | Mean I-I Corr | KR20 | Mean I-I Corr | KR20 | Mean I-I Corr | KR20 | Mean I-I Corr |
| Sequence (6 items) | 0.69 | 0.27 | 0.67 | 0.26 | 0.62 | 0.21 | 0.68 | 0.27 |
| Simple loop (5 items) | 0.72 | 0.34 | 0.71 | 0.33 | 0.66 | 0.28 | 0.72 | 0.34 |
| Nested loop (7 items) | 0.73 | 0.28 | 0.73 | 0.28 | 0.63 | 0.20 | 0.74 | 0.29 |
| If-then (2 items) | 0.63 | 0.46 | 0.61 | 0.44 | 0.70 | 0.53 | 0.66 | 0.49 |
| If-then-else (2 items) | 0.58 | 0.41 | 0.51 | 0.34 | 0.49 | 0.32 | 0.53 | 0.36 |
| While (3 items) | 0.54 | 0.28 | 0.50 | 0.25 | 0.59 | 0.32 | 0.58 | 0.32 |
| Entire scale (25 items) | 0.89 | 0.25 | 0.88 | 0.23 | 0.80 | 0.14 | 0.88 | 0.23 |

same classes, one for each grade, at two-time points, 14 to 18 days apart. For Grade 1 ($N=19$), the single measures intraclass correlation coefficient revealed a high degree of consistency among the single measures, yielding a value of 0.804 (95% CI [0.559, 0.919], F (18, 18)=9.183, $p<.001$). Similarly, Grade 2 ($N=21$) exhibited an ICC for single measures of 0.815 (95% CI [0.598, 0.921], F (20, 20)=9.809, $p<.001$). Grade 3 ($N=24$) displayed an ICC for single measures of 0.728 (95% CI [0.466, 0.872], F (23, 23)=6.353, $p<.001$). Finally, with all grades combined, the ICC for single measures remained consistent at 0.728 (95% CI [0.466, 0.872], F (23, 23)=6.353, $p<.001$). These findings collectively highlight the high degree of agreement between the single measures of the scale's responses across all grade levels.

## 6 Discussion

Our results provided a comprehensive understanding of the scale's psychometric properties in assessing primary school students' development of CT concepts such as sequences, loops, and conditionals. More specifically, the results from the validity analysis showed strong content validity, supported by high agreement among experts regarding the relevance of the scale's items to the underlying construct of CT. This suggests that the Greek scale can effectively capture the essential concepts of CT for primary school students. Furthermore, the construct validity analysis showed that our scale aligns well with the theoretical model, as evidenced by acceptable fit indices for students in Grades 1, 2, and 3 combined. Regarding construct validity for Grade 3, the results reveal a less optimal fit according to various fit indices compared to Grades 1 and 2. At the same time, we excluded Grade 4 students, as they achieved a ceiling effect. The discrepancy between the Grade 3 students and those of 1 and 2 proposes further investigation and potential adaptations of the assessment tool for this specific grade.

The analysis of the difficulty and discrimination index of the scale's items revealed that the items have varying levels of difficulty and, overall, the scale has medium difficulty. The items, however, tend to be easier for Grade 3 students. Grade 1 and Grade 2 exhibit relatively balanced difficulty levels, with Grade 2 demonstrating slightly lower overall difficulty than Grade 1. The discrimination indices also highlight stronger discriminatory power in Grade 1 and Grade 2 compared to Grade 3. These findings suggest a targeted refinement or development of items to better capture the evolving cognitive skills of Grade 3 students. The internal consistency analysis using KR20 coefficients showed high internal consistency of the overall scale across all grades and for each. However, the value of the coefficients showed that some subscales could be captured with greater accuracy than others. The latter suggests that there is the possibility of further refinement in the scale's structure. The individual grade-level analyses provide additional evidence of the validity of the assessment tool for each grade. Finally, the evaluation of the test-retest reliability indicated a high degree of consistency in all grade-level students' responses throughout 2 to 3 weeks. These finding also suggests that the scores remain consistent over time, reinforcing the scale's overall reliability.

Our results are consistent with Zapata-Cáceres et al.'s (2020, 2021) results, as both studies suggest that the scale may not be challenging for older students. Furthermore, both studies suggest that the scale's difficulty suits younger primary school students, although our findings showed a higher level of difficulty. Our study showed a great internal consistency that gradually declined as grades got higher, which is also consistent with the results obtained through the initial validation of the scale with Spanish students. However, the scores obtained from our study were lower than those obtained from the sample with which the initial validation was made. The observed variations in these results underscore the impact of cultural, educational, and contextual factors on assessment outcomes. Both studies suggest that the scale may be more suitable for Grades 1 and 2 of primary school.

Despite the promising results obtained, some limitations need to be taken into account. First, despite being adequate for the students from all grades combined, the sample size could have been better for Grade 3, which might have affected the model fit for that grade. Future research could address the specific challenges regarding the CT assessment of Grade 3 pupils and possible gender differences across the scale and its subscales. Additionally, the scale consisted of binary items and may be less able to detect subtle differences in the development of CT concepts. More diverse item types, such as multiple-choice questions or open-ended activities, might be incorporated into future studies.

Furthermore, the level of development regarding CT concepts and skills may not be adequately captured when being evaluated in a controlled classroom setting. Further item and subscale refining led by the current findings can improve the scale's overall psychometric quality and usefulness in educational assessment and research contexts. Finally, future research would focus on the inherent factor structure of the scale's items through EFA. At the same time, the integration of Item Response Theory analysis (IRT; Hambleton & Jones, 1993) could shed light on individual items' performance. Integrating these advanced statistical techniques would contribute to the refinement and optimization of the assessment tool.

## 7 Conclusion

In conclusion, our study aimed to translate, culturally adapt, and validate the BCTt for primary school students in Greece. Our comprehensive examination of the psychometric properties of the culturally adapted version of the BCTt in Greek revealed that the adapted scale has good potential for assessing students' CT skills. Our results provided evidence of the content validity, construct validity, and reliability of the Greek version of the scale. More specifically, our panel of experts agreed on the relevance of the items for assessing CT concepts, and the CFA revealed varying model fits across grades, with Grades 1 and 2 demonstrating better fits. Furthermore, our analysis revealed that the scale includes items of varying difficulty, has good discriminatory power, has good internal consistency, and has consistent and stable results over 2–3 weeks. Despite these promising outcomes, limitations were acknowledged, including sample size variations across grades, and primarily, the observed ceiling effect in Grade 4, which significantly impacted our analysis. Future research can delve into

the underlying factor structure of items through exploratory factor analysis and IRT. This approach will offer opportunities to better understand and address limitations in assessing CT skills among specific age groups. Finally, our study contributes to the growing body of research on CT assessment, offering insights for educational practitioners and researchers aiming to evaluate and enhance students' computational thinking skills.

**Acknowledgements** Ioannis Vourletsis is grateful for the opportunity to conduct this research as part of his postdoctoral studies at the Pedagogical Department of Primary Education, School of Humanities and Social Sciences, University of Thessaly.

**Author contribution** Author 1: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing. Author 2: Supervision, Project administration, Methodology, Resources, Writing - Review & Editing.

**Funding** No funding was received for conducting this study.

**Data availability** The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

## Declarations

**Ethics approval** The research adhered to ethical principles and guidelines, ensuring the protection of participants' privacy, confidentiality, and dignity throughout all stages of data collection and analysis.

**Consent to participate** Participants were provided with comprehensive study details, assured of confidentiality, and informed of their right to withdraw without consequences.

**Competing interests** The authors have no relevant financial or non-financial interests to disclose.

## References

Annamalai, S., Che Omar, A., & Abdul Salam, S. N. (2022). REVIEW OF COMPUTATIONAL THINKING MODELS IN VARIOUS LEARNING FIELDS. *International Journal of Education Psychology and Counseling*, *7*(48), 562–574. https://doi.org/10.35631/ijepc.748042.

Arafat, S., Chowdhury, H., Qusar, M., & Hafez, M. (2016). Cross Cultural Adaptation and Psychometric Validation of Research Instruments: A Methodological Review. *Journal of Behavioral Health*, *5*(3), 129. https://doi.org/10.5455/jbh.20160615121755.

Azevedo, J. M., Oliveira, E. P., & Beites, P. D. (2019). Using learning analytics to evaluate the quality of multiple-choice questions. *The International Journal of Information and Learning Technology*, *36*(4), 322–341. https://doi.org/10.1108/ijilt-02-2019-0023.

Babazadeh, M., & Negrini, L. (2022). How is computational thinking assessed in European K-12 education? A systematic review. *International Journal of Computer Science Education in Schools*, *5*(4), 3–19. https://doi.org/10.21585/ijcses.v5i4.138.

Barnard, J. J. (1999). Item analysis in test construction. *Advances in Measurement in Educational Research and Assessment*, 195–206. https://doi.org/10.1016/b978-008043348-6/50016-4.

Bartlett, M. S. (1951). The Effect of standardization on a $\chi^2$ approximation in factor analysis. *Biometrika*, *38*(3/4), 337–344. https://doi.org/10.2307/2332580.

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

Beck, C. T., & Gable, R. K. (2001). Ensuring content validity: An illustration of the process. *Journal of Nursing Measurement*, *9*(2), 201–215. https://doi.org/10.1891/1061-3749.9.2.201.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606. https://doi.org/10.1037/0033-2909.88.3.588.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health*, *6*. https://doi.org/10.3389/fpubh.2018.00149.

Borsa, J. C., Damásio, B. F., & Bandeira, D. R. (2012). Cross-cultural adaptation and validation of psychological instruments: Some considerations. *Paidéia (Ribeirão Preto)*, *22*, 423–432. https://doi.org/10.1590/1982-43272253201314.

Brennan, K., & 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. Vancouver, Canada. https://web.media.mit.edu/~kbrennan/files/Brennan_Resnick_AERA2012_CT.pdf.

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality*, *54*(1), 106–148. https://doi.org/10.1111/j.1467-6494.1986.tb00391.x.

Broglio, S. P., Ferrara, M. S., Macciocchi, S. N., Baumgartner, T. A., & Elliott, R. (2007). Test-retest reliability of computerized concussion assessment programs. *Journal of Athletic Training*, *42*(4), 509–514.

Byrne, B. M. (1994). *Structural equation modelling with EQS and EQS/Windows: Basic concepts, applications, and Programming*. Sage.

Cheung, A. K. L. (2014). Probability proportional sampling. In A. C. Michalos (Ed.), *Encyclopedia of Quality of Life and Well-Being Research* (pp. 5069–5071). Springer. https://doi.org/10.1007/978-94-007-0753-5_2269.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/bf02310555.

Cutumisu, M., Adams, C., & 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.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice-Hall.

El-Hamamsy, L., Zapata-Cáceres, M., Marcelino, P., Bruno, B., Dehler Zufferey, J., Martín-Barroso, E., & Román-González, M. (2022). Comparing the psychometric properties of two primary school Computational Thinking (CT) assessments for grades 3 and 4: The Beginners' CT test (BCTt) and the competent CT test (cCTt). *Frontiers in Psychology*, *13*(1082659). https://doi.org/10.3389/fpsyg.2022.1082659.

Epstein, J., Santo, R. M., & Guillemin, F. (2015). A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus. *Journal of Clinical Epidemiology*, *68*(4), 435–441. https://doi.org/10.1016/j.jclinepi.2014.11.021.

Fabrigar, L. R., MacCallum, R. C., Wegener, D. T., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272.

Grover, S., & Pea, R. (2018). Computational thinking: A competency whose time has come. In S. Sentence, E. Barendsen, & C. Schulte (Eds.), *Computer Science Education: Perspectives on teaching and learning* (pp. 19–38). Bloomsbury. https://doi.org/10.5040/9781350057142.ch-003.

Guggemos, J., Seufert, S., & Román-González, M. (2023). Computational thinking assessment – towards more vivid interpretations. *Technology Knowledge and Learning*, *28*(2), 539–568. https://doi.org/10.1007/s10758-021-09587-2

Hair, J. J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate Data Analysis*. Pearson Education.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x.

Hazzan, O., Ragonis, N., & Lapidot, T. (2020). Computational thinking. In O. Hazzan, N. Ragonis, & T. Lapidot (Eds.), Guide to Teaching Computer Science (pp. 57–74). Springer. https://doi.org/10.1007/978-3-030-39360-1_4.

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. *Journal of the Pakistan Medical Association, 62*(2), 142–147.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118.

Hurley, A. E., Scandura, T. A., Schriesheim, C. A., Brannick, M. T., Seers, A., Vandenberg, R. J., & Williams, L. J. (1997). Exploratory and confirmatory factor analysis: Guidelines, issues, and alternatives. *Journal of Organizational Behavior, 18*(6), 667–683. https://doi.org/10.1155/2016/2696019.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*(1), 31–36. https://doi.org/10.1007/BF02291575.

Kline, R. B. (2011). *Principles and practice of structural equation modeling, structural equation modeling*. Guilford Press.

Koo, T. K., & Li, M. Y. (2016). A Guideline of selecting and reporting Intraclass correlation coefficients for Reliability Research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151–160. https://doi.org/10.1007/BF02288391.

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

Lynn, M. R. (1986). Determination and Quantification Of Content Validity. *Nursing Research, 35*(6), 382–386. https://doi.org/10.1097/00006199-198611000-00017.

Martuza, V. R. (1977). *Applying norm-referenced and criterion-referenced measurement in education*. Allyn and Bacon.

Mitra, N. K., Nagaraja, H. S., Ponnudurai, G., & Judson, J. P. (2009). The levels of Difficulty and discrimination indices in type a multiple choice questions of pre-clinical semester 1 Multidisciplinary Summative tests. *International E-Journal of Science Medicine & Education, 3*(1), 2–7. https://doi.org/10.56026/imu.3.1.2.

Nitko, A. J., & Brookhart, S. M. (2014). Educational assessment of students (6th international electronic edition). *Harlow: Pearson*.

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

Passos, M. P. V. D., Almeida, J. R., Santos, Y. H. S., Junior, E. P. P., Flores-Quispe, M. del, Aquino, P., Martufi, R., Barreto, V., M., & Amorim, L. (2023). D. A. F. Measurement Models with Binary Indicators: A Tutorial for the Assessment of Antenatal Care Quality. https://doi.org/10.21203/rs.3.rs-2860527/v1.

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*(5), 489–497. https://doi.org/10.1002/nur.20147.

Polya, G. (1945). *How to solve it*. Princeton University Press. https://doi.org/10.1515/9781400828678.

Poulakis, E., & Politis, P. (2021). Computational thinking Assessment: Literature Review. In T. Tsiatsos, S. Demetriadis, A. Mikropoulos, & V. Dagdilelis (Eds.), *Research on E-Learning and ICT in Education*. Springer. https://doi.org/10.1007/978-3-030-64363-8_7.

Resnick, M., & Rusk, N. (2020). Coding at a crossroads. *Communications of the ACM, 63*(11), 120–127. https://doi.org/10.1145/3375546.

Revelle, W. (2021). *Psych: Procedures for Psychological, Psychometric, and Personality Research. R package version 2.2.9*. https://cran.r-project.org/package=psych.

Román-González, M. (2015). Computational thinking test: Design guidelines and content validation. In EDULEARN15 Proceedings (pp. 2436–2444). IATED. https://library.iated.org/view/ROMANGONZALEZ2015COM.

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

Román-González, M., Moreno-León, J., & Robles, G. (2019). Combining Assessment Tools for a comprehensive evaluation of computational thinking interventions. *Computational Thinking Education*, 79–98. https://doi.org/10.1007/978-981-13-6528-7_6.

Rosseel, Y. (2012). Lavaan: AnRPackage for Structural equation modeling. *Journal of Statistical Software, 48*(2). https://doi.org/10.18637/jss.v048.i02.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A review. *The Journal of Educational Research, 99*(6), 323–338. https://doi.org/10.3200/joer.99.6.323-338.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling*, Second edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Shi, J., Mo, X., & Sun, Z. (2012). Content validity index in scale development. *Journal of Central South University Medical Sciences, 37*(2), 152–155. https://doi.org/10.3969/j.issn.1672-7347.2012.02.007.

Singh, S. (2003). Simple Random Sampling. *Advanced Sampling Theory with Applications* (pp. 71–136). Springer. https://doi.org/10.1007/978-94-007-0789-4_2.

Sousa, V. D., & Rojjanasrirat, W. (2010). Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: A clear and user-friendly guideline. *Journal of Evaluation in Clinical Practice, 17*(2), 268–274. https://doi.org/10.1111/j.1365-2753.2010.01434.x.

Streiner, D. L. (2003). Starting at the beginning: An introduction to Coefficient Alpha and Internal consistency. *Journal of Personality Assessment, 80*(1), 99–103. https://doi.org/10.1207/s15327752jpa8001_18.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use*. Oxford University Press.

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

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., Bouter, L. M., & de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). *Measurement and evaluation in psychology and education* (5th ed.). Macmillan Publishing Co, Inc.

Tissenbaum, M., Sheldon, J., & Abelson, H. (2019). From computational thinking to computational action. *Communications of the ACM, 62*(3), 34–36. https://doi.org/10.1145/3265747.

Tsang, S., Royse, C., & Terkawi, A. (2017). Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. *Saudi Journal of Anaesthesia, 11*(5), 80. https://doi.org/10.4103/sja.sja_203_17.

Vaz, S., Falkmer, T., Passmore, A. E., Parsons, R., & Andreou, P. (2013). The case for using the repeatability coefficient when calculating test–retest reliability. *Plos One, 8*(9), e73990. https://doi.org/10.1371/journal.pone.0073990.

Wheaton, B., Muthén, B., Alwin, D. F., & Summers, G. F. (1977). Assessing Reliability and Stability in Panel models. *Sociological Methodology, 8*, 84–136. https://doi.org/10.2307/270754.

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* Retrieved from https://www.cs.cmu.edu/link/research-notebook-computational-thinking-what-and-why.

Yu, C. Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. University of California, Los Angeles. http://www.statmodel.com/download/Yudissertation.pdf.

Yusoff, M. S. B. (2019). ABC of Content Validation and Content Validity Index calculation. *Education in Medicine Journal, 11*(2), 49–54. https://doi.org/10.21315/eimj2019.11.2.6.

Zapata-Cáceres, M., Martín-Barroso, E., & Román-González, M. (2020). Computational thinking test for beginners: Design and content validation. *2020 IEEE Global Engineering Education Conference (EDUCON)*. https://doi.org/10.1109/educon45650.2020.9125368.

Zapata-Cáceres, M., Martín-Barroso, E., & Román-González, M. (2021). BCTt: Beginners Computational Thinking Test. In Understanding computing education (Vol 1). Proceedings of the Raspberry Pi Foundation Research Seminar series. Retrieved from www.rpf.io/seminar-proceedings-2020.

Zapata-Cáceres, M., Marcelino, P., El-Hamamsy, L., & Martín-Barroso, E. (2024). A Bebras Computational thinking (ABC-Thinking) program for primary school: Evaluation using the competent computational thinking test. *Education and Information Technologies*. https://doi.org/10.1007/s10639-023-12441-w.