Which cognitive abilities underlie computational thinking? Criterion validity of the Computational Thinking Test

Marcos Román-González*, Juan-Carlos Pérez-González, Carmen Jiménez-Fernández

Universidad Nacional de Educación a Distancia (UNED), Faculty of Education, C/ Juan del Rosal, nº 14, C.P. 28040, Madrid, Spain

A R T I C L E I N P R E S S

Computers in Human Behavior xxx (2016) 1–14

1. Introduction

We live immersed in a digital ecosystem full of objects driven by software (Manovich, 2013). In this context, being able to handle the language of computers is emerging as an inescapable skill, a new literacy, which allows us to participate fully and effectively in the world of digital objects driven by software. However, there is still no consensus on a CT definition or how to measure it. In response, we attempt to address both issues from a psychometric approach. On the one hand, a Computational Thinking Test (CTT) is administered on a sample of 2,515 Spanish students from 5th to 10th grade, so its descriptive statistics and reliability are reported in this paper. On the second hand, the criterion validity of the CTT is studied with respect to other standardized psychological tests: the Primary Mental Abilities (PMA) battery, and the RPM problem-solving test. Thus, it is intended to provide a new instrument for CT measurement and additionally give evidence of the nature of CT through its associations with key related psychological constructs. Results show statistically significant correlations at least moderately intense between CT and: spatial ability (r = 0.44), reasoning ability (r = 0.44), and problem-solving ability (r = 0.67). These results are consistent with recent theoretical proposals linking CT to some components of the Cattell-Horn-Carroll (CHC) model of intelligence, and corroborate the conceptualization of CT as a problem-solving ability.

© 2016 Elsevier Ltd. All rights reserved.

Keywords: Computational thinking, Computational Thinking Test, Code literacy, Cognitive abilities, Computer science education, Educational psychology, Primary education, Secondary education

ABSTRACT

Computational thinking (CT) is being located at the focus of educational innovation, as a set of problem-solving skills that must be acquired by the new generations of students to thrive in a digital world full of objects driven by software. However, there is still no consensus on a CT definition or how to measure it. In response, we attempt to address both issues from a psychometric approach. On the one hand, a Computational Thinking Test (CTT) is administered on a sample of 2,515 Spanish students from 5th to 10th grade, so its descriptive statistics and reliability are reported in this paper. On the second hand, the criterion validity of the CTT is studied with respect to other standardized psychological tests: the Primary Mental Abilities (PMA) battery, and the RPM problem-solving test. Thus, it is intended to provide a new instrument for CT measurement and additionally give evidence of the nature of CT through its associations with key related psychological constructs. Results show statistically significant correlations at least moderately intense between CT and: spatial ability (r = 0.44), reasoning ability (r = 0.44), and problem-solving ability (r = 0.67). These results are consistent with recent theoretical proposals linking CT to some components of the Cattell-Horn-Carroll (CHC) model of intelligence, and corroborate the conceptualization of CT as a problem-solving ability.

* Corresponding author.

E-mail addresses: mroman@edu.uned.es (M. Román-González), jcperez@edu.uned.es (J.-C. Pérez-González), mjimenez@edu.uned.es (C. Jiménez-Fernández).

http://dx.doi.org/10.1016/j.chb.2016.08.047
0747-5632/© 2016 Elsevier Ltd. All rights reserved.
However, there is still little consensus on a formal definition of CT (Gouws, Bradshaw, & Wentworth, 2013; Kalelioglu, Gulbahar, & Kukul, 2016), and disagreements over how it should be integrated in educational curricula (Lye & Koh, 2014). Similarly, there is a worrying vacuum about how to measure and assess CT, fact that must be addressed. Without attention to assessment, CT can have little hope of making its way successfully into any curriculum. Furthermore, in order to judge the effectiveness of any curriculum incorporating CT, measures that would enable educators to assess what the student has learned need to be validated (Grover & Pea, 2013).

In response, we attempt to address these issues from a psychometric approach. On the one hand, how our Computational Thinking Test (CTt) has been designed and developed is reported, as well as its descriptive statistics and reliability derived from an administration on a sample exceeding a thousand Spanish students. On the other hand, the criterion validity (Cronbach & Meehl, 1955) of the CTt is studied with respect to already standardized psychological tests of core cognitive abilities. Thus, this paper is aimed at providing a new instrument for measuring CT and additionally giving evidence of the correlations between CT and other well-established psychological constructs in the study of cognitive abilities.

1.1. Computational thinking definitions

We can distinguish between: a) generic definitions; b) operational definitions; c) educational and curricular definitions.

1.1.1. Generic definitions

One decade ago, in 2006, Jeanette Wing’s foundational paper defined that CT “involves solving problems, designing systems, and understanding human behavior, by drawing on the concepts fundamental to computer science” (Wing, 2006, p. 33). Thus, CT’s essence is thinking like a computer scientist when confronted with a problem. But this first generic definition has been revisited and specified in successive attempts over the last few years, still not reaching an agreement (Grover & Pea, 2013; Kalelioglu et al., 2016). So, in 2011 Wing clarified, CT “is the thought processes involved in formulating problems and their solutions so that the solutions are represented in a form that can be effectively carried out by an information-processing agent” (Wing, 2011; on-line). One year later, this definition is simplified by Aho, who conceptualizes CT as the thought processes involved in formulating problems so “their solutions can be represented as computational steps and algorithms” (Aho, 2012, p. 832).

1.1.2. Operational definitions

In 2011, the Computer Science Teachers Association (CSTA) and the International Society for Technology in Education (ISTE) developed an operational definition of computational thinking that provides a framework and common vocabulary for Computer Science K-12 educators: CT is a “problem-solving process that includes (but is not limited to) the following characteristics: formulating problems in a way that enables us to use a computer and other tools to help solve them; logically organizing and analyzing data; representing data through abstractions such as models and simulations; automating solutions through algorithmic thinking (a series of ordered steps); identifying, analyzing, and implementing possible solutions with the goal of achieving the most efficient and effective combination of steps and resources; generalizing and transferring this problem solving process to a wide variety of problems” (CSTA & ISTE, 2011; on-line).

1.1.3. Educational-curricular definitions

More than definitions in the strict sense, frameworks for developing CT in the classroom and other educational settings are mentioned next. So, from the UK, the organization Computing At School (CAS) states that CT involves six different concepts (logic, algorithms, decomposition, patterns, abstraction, and evaluation), and five approaches to working (tinkering, creating, debugging, persevering, and collaborating) in the classroom (CAS Barefoot, 2014). Moreover, from the United States, Brennan and Resnick (2012) describe a CT framework that involves three key dimensions: “computational concepts” (sequences, loops, events, parallelism, conditionals, operators, and data); ‘computational practices’ (experimenting and iterating, testing and debugging, reusing and remixing, abstracting and modularizing); and ‘computational perspectives’ (expressing, connecting, and questioning). Table 1 shows a crosstab intersecting the CT framework dimensions (Brennan & Resnick, 2012) with the sampling domain of our Computational Thinking Test (CTt), which will be detailed in Sub-section 1.4.

1.2. Computational thinking from the CHC model of intelligence

While CT involves thinking skills to solve problems algorithmically (e.g., Brennan & Resnick, 2012; Grover & Pea, 2013), intelligence (i.e., general mental ability or general cognitive ability) involves primarily the ability to reason, plan and solve problems (Gottfredson, 1997). Even authors with alternative approaches to the conceptualization of intelligence recognize intelligence as a “computational capacity” or “the ability to process certain kinds of information in the process of solving problems of fashioning products” (Gardner, 2006, p. 503).

Within a cognitive approach, it has been recently suggested (Ambrosio, Xavier, & Georges, 2014) that computational thinking is related to the following three abilities-factors from the Cattell-Horn-Carroll (CHC) model of intelligence (McGrew, 2009; Schneider & McGrew, 2012):

- **Fluid reasoning** ($G_F$), defined as: “the use of deliberate and controlled mental operations to solve novel problems that cannot be performed automatically. Mental operations often include drawing inferences, concept formation, classification, generating and testing hypothesis, identifying relations, comprehending implications, problem solving, extrapolating, and transforming information. Inductive and deductive reasoning are generally considered the hallmark indicators of $G_F$” (McGrew, 2009, p. 5).

- **Visual processing** ($G_V$), defined as “the ability to generate, store, retrieve, and transform visual images and sensations. $G_V$ abilities are typically measured by tasks (figural or geometric stimuli) that require the perception and transformation of visual shapes, forms, or images and/or tasks that require maintaining spatial orientation with regard to objects that may change or move through space” (McGrew, 2009, p. 5).

- **Short-term memory** ($G_m$), defined as “the ability to apprehend and maintain awareness of a limited number of elements of information in the immediate situation (events that occurred in the last minute or so). A limited-capacity system that loses information quickly through the decay of memory traces, unless an individual activates other cognitive resources to maintain the information in immediate awareness” (McGrew, 2009, p. 5).

Therefore, it is expected that a computational thinking test should correlate with other already validated tests aimed at measuring cognitive abilities cited above.
1.3. Computational thinking assessment

Count on validated measurement instruments is something necessary and valuable in any research area. However, for the moment, there is still a large gap in tests relating to CT that have undergone a comprehensive psychometric validation process (Mühling, Ruf, & Hubwieser, 2015). As Buffum et al. (2015) say: “developing assessments of student learning is an urgent area of need for the relatively young computer science education community as it advances toward the ranks of more mature disciplines such as physics that have established standardized assessments over time” (Buffum et al., 2015, p. 622). Anyway, we find in recent years some remarkable attempts to measure and assess CT in students from 5th to 10th grade, which are the ones of this paper’s interest.

From the University of California, comes the instrument Fairy Assessment in Alice (Werner, Denner, Campe, & Kawamoto, 2012), which tries to measure the understanding and use of abstraction, conditional logic, algorithmic thinking and other CT concepts that middle school students utilize to solve problems. However, this instrument is designed ad hoc to be used in the context of programming learning environment Alice1 (Graczyńska, 2010), and it has not been undergone to a psychometric validation process. The research group from Clemson University (South Carolina) provides a complementary perspective (Daily, Leonard, Jörg, Babu, & Gundersen, 2014; Leonard et al., 2015). These authors propose a kinesthetic approach to learning (‘embodied learning’) and assessment of CT with 5th and 6th grade students. To do so, they alternate activities for programming motion sequences (choreographies) in the Alice environment, with the representation of those same sequences in a physical-kinesthetic environment. The assessment tool also combines both settings, but its psychometric properties have not been reported.

Another interesting research line with middle school students is provided by the group from the University of Colorado. They work with students in the video-game programming environment AgentSheets2. Within a first group of studies (Koh, Basawapatna, Bennett, & Repenning, 2010), these authors identify several Computational Thinking Patterns (CTP) that young programmers abstract and develop during the creation of their video-games; in this context, they design the Computational Thinking Patterns Graph, an automated tool that analyzes the games programmed by the students, and represents graphically how far each game has involved the different CTP when compared with a model. Within a second group of studies (Basawapatna, Koh, Repenning, Webb, & Marshall, 2011), the authors try to assess whether students are able to transfer the CTP acquired during video-game programming to a new context of scientific simulations programming. For this assessment, they develop CTP-Quiz instrument, whose reliability or validity have not been reported.

Similarly, from the Universidad Rey Juan Carlos (Madrid, Spain) Dr. Scratch3 is presented (Moreno-León & Robles, 2015a, 2015b, 2014). Dr. Scratch is a free and open source web application designed to analyze, simply and automatically, projects programmed with Scratch4 (Resnick et al., 2009), as well as it provides feedback that can be used to improve programming skills and to develop CT in middle school students (Moreno-León, Robles, & Román-González, 2015). In order to assign an overall CT score to the project, Dr. Scratch infers the programmer competence along the following seven CT dimensions: Abstraction and problem decomposition; Parallelism; Logical thinking; Synchronization; Flow control; User Interactivity; and Data representation. Therefore, Dr. Scratch is not strictly a cognitive test but a tool for the formative assessment of Scratch projects. Dr. Scratch is currently under validation process, although its convergent validity with respect to other traditional metrics of software quality and complexity has been already reported (Moreno-León, Robles, & Román-González, 2016).

Furthermore, we consider the Bebras International Contest,5 a competition born in Lithuania in 2003 which aims to promote the interest and excellence of primary and secondary students around the world in the field of Computer Science from a CT perspective (Cartell, Dagiene, & Futschek, 2012; Dagiene & Futschek, 2008; Dagiene & Stupurienė, 2014). Each year, the contest proposes a set of Bebras Tasks, whose overall approach is the resolution of real problems, significant for the students, through the transfer and projection of their CT over those. These Bebras Tasks are independent from any particular software or hardware, and can be administered to individuals without any prior programming experience.

---

1 http://www.alice.org/index.php
2 http://www.agentsheets.com/
3 http://drscratch.org/
4 https://scratch.mit.edu/
5 http://www.bebras.org/

---

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Description</th>
<th>Components</th>
<th>CT framework</th>
<th>Sampling domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computational concepts</td>
<td>Concepts students employ as they program</td>
<td>Sequences, Loops, Events, Parallelism, Conditionals, Operators, Data</td>
<td>*</td>
<td>Computational concept addressed</td>
</tr>
<tr>
<td>Computational practices</td>
<td>Problem-solving practices that occurs in the process of programming</td>
<td>Experimenting and iterating, Testing and debugging, Reusing and remixing, Abstracting and modularizing, Expressing, Connecting, Questioning</td>
<td>/</td>
<td>Required task</td>
</tr>
<tr>
<td>Computational perspectives</td>
<td>Students’ understandings of themselves, their relationships to others, and the digital world around them</td>
<td></td>
<td>/</td>
<td></td>
</tr>
</tbody>
</table>
For all these features, the Bebras Tasks have been pointed out as more than likely embryo for a future PISA (Programme for International Student Assessment) test in the field of Computer Science (Hubwieser & Mühlung, 2014; Jašková & Kováčová, 2015). Anyway, the Bebras International Contest is, at the moment, an event for promoting CT, not a measuring instrument; among other considerations, because it is not composed by a stable and determined set of task-items, but a set that varies from year to year, with slight modifications along the countries. However, its growing expansion has aroused the interest of psychometry researchers, who have begun to investigate its possible virtues as a CT measurement instrument. Thus, descriptive studies about the student’s performance on Bebras Tasks have been recently published, referred to the corresponding editions of the Bebras International Contest held in Germany (Hubwieser & Mühlung, 2014, 2015), Italy (Bellettini et al., 2015), Taiwan (Lee, Lin, & Lin, 2014) or Turkey (Kalelioğlu, Gülbahar, & Madran, 2015). In all of them, and in most of the tasks studied, significantly higher performances in the male group in comparison with the female group were reported.

But strictly speaking, we only have knowledge of two tests aimed to middle/high school students which are being fully subjected to the psychometric requirements; both instruments are currently undergoing a validation process.

a. Test for Measuring Basic Programming Abilities (Mühlung et al., 2015): it is designed for Bavarian students from 7th to 10th grade. This test is aimed at measuring the students’ ability to execute a given program based on the so-called ‘flow control structures’; which are considered at the core of the CT for this age group: Sequencing (doing one step after another); Selection (doing either one thing or another); Repetition (doing one thing once and again). These control structures lead to the following CT concepts that are covered by the test: sequence of operations; conditional statement with (if/else) and without (if) alternative; loop with fixed number of iterations (repeat times); loop with exit condition (conditional loop: while or repeat until); and the nesting of these structures to create more complex programs.

b. Commutative Assessment (Weintrop & Wilensky, 2015): it is designed for high-school students, from 9th to 12th grade. This test is aimed at measuring students’ understanding of different computational concepts, depending on whether they occur through scripts written in visual (block-based) or textual programming languages; which is a key transition to reach higher levels of code-literacy. The test has a length of 28 items, and it addresses the following CT concepts: conditionals; defined/fixed loops; undefined/unfixed loops; simple functions; functions with parameters/variables.

1.4. Computational Thinking Test

Overall, our Computational Thinking Test (CTt) has been developed following the practical guide to validating computer science knowledge assessments with application to middle school from Buffum et al. (2015), which is aligned with the international standards for psychological and educational testing (AERA, APA, & NCME, 2014). In addition, the CTt is consistent with other computational thinking tests under validation, aimed to middle/high school, such as the Test for Measuring Basic Programming Abilities (Mühlung et al., 2015) or the Commutative Assessment (Weintrop & Wilensky, 2015), just described in Sub-section 1.3.

The CTt was initially designed with a length of 40 multiple choice items (version 1.0, October 2014). After a content validation process through twenty experts’ judgement, this first version was refined to the final one (version 2.0, December 2014) of 28 items length (Román-González, 2015); which is built on the following principles:

- **Aim:** CTt aims to measure the development level of CT in the subject.
- **Operational definition of measured construct:** CT involves the ability to formulate and solve problems by relying on the fundamental concepts of computing, and using logic-syntax of programming languages: basic sequences, loops, iteration, conditionals, functions and variables.
- **Target population:** CTt is mainly designed and intended for Spanish students between 12 and 14 years old (7th and 8th grade); although it can be also used in lower grades (5th and 6th grade) and upper grades (9th and 10th grade).
- **Instrument Type:** multiple choice test with 4 answer options (only one correct).
- **Length and estimated completion time:** 28 items; 45 min.

Each item of the CTt is designed and characterized according to the following five dimensions of the sampling domain:

- **Computational concept addressed:** each item addresses one or more of the following seven computational concepts, ordered in increasing difficulty: Basic directions and sequences (4 items); Loops—repeat times (4 items); Loops—repeat until (4 items); If—simple conditional (4 items); If/else—complex conditional (4 items); While conditional (4 items); Simple functions (4 items). These ‘computational concepts’ are aligned with some of the CT framework (Brennan & Resnick, 2012; see Table 1) and with the CSTA Computer Science Standards for 7th and 8th grade (CSTA, 2011).
- **Environment-Interface of the item:** CT items are presented in any of the following two environments-interfaces: ‘The Maze’ (23 items) or ‘The Canvas’ (5 items). Both interfaces are common in popular sites for learning programming such as Code.org (Kalelioğlu, 2015).
- **Answer alternatives style:** in each item, the response alternatives may be presented in any of these two styles: Visual arrows (8 items) or Visual blocks (20 items). Both styles are also common in popular sites for learning programming such as Code.org (Kalelioğlu, 2015).
- **Existence or non-existence of nesting:** depending on whether the item solution involves a script with (19 items) or without (9 items) nesting computational concepts (a concept embedded in another to a higher hierarchy level) (Mühlung et al., 2015).
- **Required task:** depending on which of the following cognitive tasks is required for solving the item: Sequencing: the student must sequence, stating in an orderly manner, a set of commands (14 items); Completion: the student must complete an incomplete given set of commands (9 items); Debugging: the student must debug an incorrect given set of commands (5 items). This dimension is partially aligned with the aforementioned ‘computational practices’ from the CT framework (Brennan & Resnick, 2012; see Table 1).

The CTt is administered collectively and on-line, and it can be performed both via non-mobile or mobile electronic devices. Preliminary results about the CTt psychometric properties after its administration on a sample of 400 Spanish students (7th and 8th grade) have been already reported (Román-González, Pérez-
2. Method

2.1. Participants

The CTt was administered on a total sample of 1,251 Spanish students, boys and girls from 24 different schools enrolled from 5th to 10th grade. The distribution of the subjects by gender, grade and age is shown in Table 2. From the total sample, 825 (65.9%) students belong to public schools, and 426 (34.1%) belong to private schools. Considering the device on which the CTt was administered, 1,001 students did it on a personal computer (80.0%) and 250 students (20.0%) did it so on a tablet. None of the subjects had prior programming formal experience when the CTt was administered.

The sampling procedure is not probabilistic and intentional. Depending on the reasons that led to sample the different subjects, these can be divided into four sub-samples:

- Sub-sample A (n = 418): it is composed of individuals belonging to classrooms that subsequently enrolled in the Accelerated Intro to CS Course from Code.org.
- Sub-sample B (n = 48): it is composed of individuals belonging to classrooms that subsequently enrolled in the CS Fundamentals Course 2 from Code.org.
- Sub-sample C (n = 194): it is composed of individuals belonging to classrooms that subsequently started to learn programming with Scratch.
- Sub-sample D (n = 591): it is composed of individuals belonging to classrooms that, although they did not subsequently start to learn programming, were interested in measuring the CT of the students.

In addition to our CTt, other standardized tests were administered concurrently to a part of the above subjects. Specifically for this paper, administrations of Primary Mental Abilities (PMA) battery, and the RP30 problem-solving test; which are described next.

2.2. Instruments

In order to address the criterion validation of the CTt, another two standardized instruments are administered: the Primary Mental Abilities (PMA) battery, and the RP30 problem-solving test, which are described next.

2.2.1. Primary Mental Abilities (PMA) battery

The PMA battery is aimed at appreciating the basic cognitive abilities through four different subtests, which allow an estimate of the main components of intelligence. This is a well-known measure of cognitive abilities (e.g., Hertzog & Bleckley, 2001; Quiroga et al., 2015) developed by Thurstone (1938). Its maximum administration time is 26 min, and it can be used from 10 years old onwards. The Spanish technical manual (TEA Ediciones, 2007) reports excellent reliability and validity coefficients about the four subtests. The PMA provides a precise measurement of the following cognitive abilities:

- **Verbal factor (PMA-V):** Ability to understand and express ideas with words. PMA-V items involve selecting the accurate synonym of a word given.
- **Spatial factor (PMA-S):** Ability to imagine and devise objects in two and three dimensions. PMA-S items involve selecting equal figures to a given model, after having been rotated.
- **Reasoning factor (PMA-R):** Ability to solve logical problems, to understand and plan. PMA-R items involve selecting the option which continues a logical series given.
- **Numerical factor (PMA-N):** Ability to handle numbers and quantitative concepts. PMA-N items involve checking mentally the sum of four two-digit numbers.

2.2.2. RP30 problem-solving test

RP30 problem-solving test is aimed to assess speed and flexibility in performing logical operations. Its maximum administration time is 17 min, and it can be used from 12 years old onwards. The Spanish technical manual (Seisdedos, 2002) reports excellent reliability values for RP30 ($r_{xx} > 0.90$; through the split-half method), as well as its criterion validity regarding to Changes Test of Cognitive Flexibility ($r_{xy} = 0.38$) or to DAT$^{10}$-Spatial ($r_{xy} = 0.34$).

Fig. 1. CTt, item 6: loops–repeat times; 'The Maze'; visual arrows; no-nesting; completion.

---

9 Test Cambios de Flexibilidad Cognitiva [Changes Test of Cognitive Flexibility] (Seisdedos, 1994).
10 DAT: Differential Aptitude Tests (Bennett, 1952).
Fig. 2. CTt, item 7: loops—repeat times; ‘The Canvas’; visual blocks; no-nesting; debugging.

Fig. 3. CTt, item 14: loops—repeat until + If—simple conditional; ‘The Maze’; visual blocks; yes-nesting; sequencing.

Fig. 4. CTt, item 25: loops—repeat times + simple functions; ‘The Canvas’; visual blocks; yes-nesting; sequencing.
RP30 appreciates a high-level cognitive ability, by which a series of logical relations given are understood by the subject in order to determine whether these relations are met in several simple conditions. The subject must decide whether the problem contains logical relationships, which allow the subject to retain the given information. RP30 requires enough concentration as errors are penalized. It is considered that there are three cognitive abilities underlying RP30 performance (Seisdedos, 2002):

- **Reasoning**, due to the fact that the logical relations which may be satisfied by the subject must be previously understood by the subject.
- **Spatial ability**, as the subject must process the small circles and squares contained in each item, in order to decide if the condition is satisfied.
- **Working memory**, which allows the subject to retain the given logical relation without need of constantly consulting it.

### 2.3. Procedure

Participating subjects in our research were enrolled in the elective subject of Computer Science, which is held twice a week (1 h each). Typically, the CTt was administered during the first of the two weekly classes. In the groups in which another standardized instrument was further administered, it was done during the second weekly class.

For the CTt collective administration, the Computer Science teacher followed the instructions which were sent by email in the week before, containing the URL to access the on-line test. The student’s direct answers to the CTt items were stored in the Google Drive database linked with the instrument, which was subsequently downloaded as an Excel.xls file.

For the collective administration of the standardized instrument (PMA or RP30), students were previously signed in the on-line platform from the publishing house,\(^\text{11}\) holder of these tests’ commercial rights. Come the administration day, the subjects logged in the platform and performed the corresponding instrument, PMA battery or RP30 test (never both). Afterwards, from our administrator profile, we could download the subjects’ results as an Excel.xls file.

Finally, all .xls files generated during data collection were exported to a single .sav file, which constitutes the data matrix under analysis with SPSS software (version 22). From this analysis arise the results exposed below.

### 3. Results and discussion

#### 3.1. Descriptive statistics

Table 3 shows the main descriptive statistics of the CTt score (calculated as the sum of correct answers along the 28 items of the test) for the entire sample (n = 1,251).

In Fig. 6 (left), a histogram showing the distribution of the CTt score along the sample is depicted. As it can be seen, the aforementioned distribution fits remarkably the normal curve; although, given the very large size of the sample, the small existing maladjustments are penalized by the Kolmogorov-Smirnov test which rejects the null hypothesis of normality ($Z_{KS} = 0.052; p < 0.01$).

In Fig. 6 (right), we show the success rate per item (expressed in per unit) or item difficulty index, that confirms empirically the progressive difficulty of the CTt; which was already anticipated by the experts during the content validation process (Román-González, 2015). The average success rate along the 28 items is $p = 0.59$ (medium difficulty); ranging from $p = 0.16$ (item 23; very high difficulty) to $p = 0.96$ (item 1; very low difficulty).

Summarizing, it can be stated that: a) the CTt score is almost normally distributed (i.e. symmetrically distributed: skewness $=0$), showing proper variability so that is possible to construct suitable scales (percentiles) for the target population; b) the CTt has an appropriate degree of difficulty (medium) for the target population, with an increasing difficulty along its items, as recommended in the design of abilities’ tests (e.g., Carpenter, Just, & Shell, 1990; Eilithorn & Telford, 1969).

#### 3.1.1. Differences by grade

When the sample is segmented regarding to grade, the descriptive statistics shown in Table 4 are obtained. Specifically, results in Table 4 are split according to the Spanish educational system by the end of Primary Education (5th and 6th grade), the start of Secondary Education (7th and 8th grade), and the end of Secondary Education (9th and 10th grade).

Box plots for the CTt score split by aforementioned grades are shown in Fig. 7. The outlier belongs to a case from 6th grade, which obtained CTt score equal to 26 (i.e., $\approx$ 3 standard deviations above the mean of its reference group). The ANOVA test shows statistically significant differences in the CTt score regarding to grade ($F_2,1248 = 50.514; p < 0.01$). The post-hoc Tukey test additionally shows statistically significant differences between all possible pairs of means ($p < 0.01$).

Hence, it can be stated that the performance on the CTt increases as it does the grade; this result is consistent with our assumption that the CT is a problem-solving ability that it should be therefore linked to the cognitive development and maturity of the subjects (Ackerman & Rolfhus, 1999; Mayer, Caruso, & Salovey, 1999).

#### 3.1.2. Gender differences

About the possible differential performance on the CTt regarding to gender, we find a statistically significant difference in the CTt score in favor of the male group ($t = 5.374; p < 0.01$), resulting an effect size measured through Cohen’s $d$ (Cohen, 1992) equal to 0.31 (Table 5); that can be considered as a low-moderate effect. If the aforementioned difference is analyzed along grades (Table 5), higher means in the CTt score are always found in the

---

\( ^{11}\) http://www.e-teaediciones.com/
male group, although these gender differences are only statistically significant from 7th and 8th grade ($t = 2.928; p < 0.01$) onwards; being more intense in 9th and 10th grade ($t = 3.451; p < 0.01$). Hence, it seems that there is a progressive gender gap over the CTt performance, as we advance along the grades (Fig. 8).

These gender differences are consistent with those found in previous research with Bebras Tasks, on which most of the investigations report higher yields of the male group, as described in Sub-section 1.3.
and subjects who did it so on a tablet: for instance, in 7th and 8th grade, \( \bar{X}_{\text{Computer}} = 16.01 \) vs. \( \bar{X}_{\text{Tablet}} = 18.24 \) (\( t = 4.116; p < 0.01; d = 0.50 \)). In the future, if we achieve a larger sample of subjects who perform the CTt on tablet, and if these aforementioned significant differences between devices continue, it will be necessary to construct different scales for the CTt depending on the administration device.

3.3. Criterion validity

3.3.1. Relative to Primary Mental Abilities (PMA) battery

Correlations between the CTt and the various tests of the PMA battery are shown in Table 7. As it can be seen, the CTt has a positive statistically significant correlation (\( p < 0.01 \)), moderately intense with PMA-R (reasoning factor) and PMA-S (spatial factor), and slightly intense with PMA-V (verbal factor). There is no statistically significant correlation between CTt and PMA-N (numerical factor). Corresponding scatter plots are shown in Fig. 9.

At this point, we perform a multiple linear regression over the CTt score (considered as the dependent variable) based on the PMA-V, PMA-S, PMA-R and PMA-N scores (considered as predictors). Table 8 summarizes de regression model, which is calculated through the ‘enter’ method. This regression model, based on the PMA battery, correlates \( r = 0.54 \) with the CTt; which means an adjusted \( R^2 = 0.27 \). That is, the 27.0% of the CTt scores’ variance is explained from a linear combination of the primary mental abilities measured through the PMA battery. Normality of the regression model residuals was verified.

The regression model is able to explain, statistically significant, the differences in the CTt scores, as \( F_{(4, 131)} = 13.457 \) (\( p < 0.01 \)). However, as shown in following Table 9 which contains the coefficients of the regression model, only PMA-S (spatial factor) and PMA-R (reasoning factor) are capable, specifically and statistically significant (\( p < 0.01 \)), to explain differences in the dependent variable (CTt). The standardized coefficients of the model are, from highest to lowest value, \( \beta_{\text{PMA-S}} = 0.308; \beta_{\text{PMA-R}} = 0.265; \beta_{\text{PMA-V}} = 0.134; \beta_{\text{PMA-N}} = -0.051 \).

From our perspective, these results point out two important issues:

- Firstly, there is still a 73.0% of the CTt scores’ variance that is not explained by the primary mental abilities measured through the PMA battery; which suggests certain independence of CT as a psychological construct, distinct from the traditional aptitudes.
- Secondly, the cognitive abilities with higher explanatory power about CT are reasoning ability and spatial ability; from both there is abundant evidence in the literature that reports certain male superiority. Regarding to the former, Kuhn and Holling (2009) recently report gender differences in reasoning ability favoring males in German students from 7th to 10th grade. Regarding to the latter, there are some meta-analysis that demonstrate higher male spatial ability, especially in tasks that involve mentally rotation of figures (Linn & Petersen, 1985; Voyer, Voyer, & Bryden, 1995). All the above could explain the higher yield of the boys in the CTt seen in Sub-section 3.1.2.

3.3.2. Relative to RP30 problem-solving test

Correlation between CTt and RP30 problem-solving test is shown in Table 10. As it can be seen, we find a positive, statistically significant, and moderately-strongly intense correlation (\( r = 0.669; p < 0.01 \)) between both instruments. Corresponding scatter plot is shown in Fig. 10, such as the coefficient of determination \( R^2 = 0.447 \) (i.e., 44.7% of shared variance between both scores). Recall that RP30 test appreciates a high-level cognitive ability and it has been previously used as a proxy of the general mental ability. Our results
indicate that CTt correlate more intensely with RP30 than with any of the primary mental abilities measured through PMA battery (Table 11). Hence, it seems that computational thinking could be fundamentally linked with general mental ability (particularly with fluid intelligence); and to a lesser extent with different cognitive aptitudes, such as logical reasoning and spatial ability.

When results of preceding Sub-sections 3.3.1 and 3.3.2 are triangulated, we find a clear consistency between the magnitude of the correlations CTt\*PMA and CTt\*RP30, and the expected composition of computational thinking from the CHC model of intelligence exposed in Sub-section 1.2 (Table 11). From our point of view, this is a powerful evidence of the criterion concurrent validity of our CTt, as well as an empirical confirmation of the computational thinking construct’s composition proposed by Ambrosio et al. (2014).

4. Implications and limitations

The CTt has some strengths like: it can be administered in pretest conditions to measure the initial development level of CT in students without prior programming experience from 5th to 10th grade; it can be collectively administered so it could be used in massive screenings and early detection of students with high abilities (or special needs) for programming tasks; it can be utilized for collecting quantitative data in pre-post evaluations of the efficacy of curricula or programs aimed at fostering CT, which would be a desirable practice versus the qualitative approach that has been mostly used in the literature so far (Lye & Koh, 2014); and it could be used along academic and professional guidance processes towards STEM disciplines. However, the CTt also has obvious limitations and weaknesses:

- The CTt provides a static and decontextualized assessment. Therefore, we recommend to complement its use with other CT assessment tools designed from a formative perspective, such as Dr. Scratch (Moreno-León et al., 2015).
- In terms of CT framework (Brennan & Resnick, 2012), the CTt is overly focused on ‘computational concepts’, only covers ‘computational practices’ partly, and ignores ‘computational perspectives’.
- The CTt only demands the projection of computational thinking over logical and visuospatial problems, such as solving mazes or designing geometric patterns. This implies a clear bias of the CTt, as computational thinking can also be projected over problems with different features, such as: modeling scientific simulations (Weintrop et al., 2016); algorithmic composition of computational music (Edwards, 2011); or digital interactive storytelling (Burke, 2012; Howland & Good, 2015). The latter authors report significantly higher values in the computational complexity of scripts written by girls from 7th and 8th grade in comparison with their male peers within narrative tasks; this
result is consistent with the (slight) female superiority in tasks involving verbal ability reported in the classical literature (Hyde & Linn, 1988). It seems, therefore, that the direction of gender differences in CT may vary depending on the type of problems on which such ability is projected.

- Finally, as the CTt is entirely designed with multiple choice items, it might be measuring CT at its lower cognitive complexity levels (‘recognize’ and ‘understand’) (Gouws et al.,...
5. Conclusions and further research

In this paper we have provided evidences of reliability and criterion validity of a new instrument for the assessment of CT and additionally we expanded our understanding of the CT nature through the theory-driven exploration of its associations with other established psychological constructs in the cognitive sphere. We have found expected positive small or moderate significant correlations (0.27 < r < 0.44) between CT and three of the four primary mental abilities of the Thurstone (1938) model of intelligence, as well as a high correlation (r = 0.67) between CT and problem-solving ability as a proxy of general mental ability. Our findings are consistent with recent theoretical proposals by Ambrosio et al. (2014) linking CT with some core elements of the CHC model of intelligence (McGrew, 2009), especially with respect to \( G_f \) (fluid intelligence) and \( G_v \) (visual processing). Furthermore, our results support the statement that CT is fundamentally linked with general mental ability; and also, to a lesser extent, with specific cognitive aptitudes, such as inductive reasoning, spatial and verbal abilities. This corroborates the conceptualization of CT as a problem-solving ability (e.g., Brennan & Resnick, 2012; Lye & Koh, 2014; Wing, 2006, 2008); and it is consistent with the framework recently described by Kalelioglu et al. (2016), in which CT is defined as a complex and high-order thinking skill involved in problem-solving processes.

Overall, it should be noted that this paper contributes to the establishment of the nomological net (Cronbach & Meehl, 1955) of computational thinking as an emergent scientific construct. Future research might expand this nomological net exploring how CT is related to other cognitive and computational variables, such as working memory, executive functions, or specific programming skills, among others. Finally, we plan the following further research lines concerning the CTt: a) convergent validity studies between CTt and other alternative CT assessment tools, such as Dr. Scratch (Moreno-León & Robles, 2015b, 2015a), Bebras Tasks (Dagiene & Stupuriene, 2014), the Test for Measuring Basic Programming Abilities (Mühling et al., 2015), or the Commutative Assessment (Weintrop & Wilensky, 2015); b) CTt adaptation and translation into other languages (already underway adaptations-translations into English and Portuguese), and replications of our psychometric studies in other populations; c) enhanced CTt versions including items that demand to project and transfer CT on scientific, narrative and musical relevant problems.

Acknowledgements

We thank Professor Dr. Kate Howland (University of Sussex) for collaborating in the adaptation and translation of CTt items from the Spanish language to the English language.

References


