



# Quantifying Computational Thinking Skills: an Exploratory Study on Bebras Tasks

Ana Liz Souto Oliveira   [ Federal University of Paraíba | [analiz@dcx.ufpb.br](mailto:analiz@dcx.ufpb.br) ]

Wilkerson L. Andrade  [Federal University of Campina Grande|[wilkerson@computacao.ufcg.edu.br](mailto:wilkerson@computacao.ufcg.edu.br) ]

Dalton Serey  [ Federal University of Campina Grande | [dalton@computacao.ufcg.edu.br](mailto:dalton@computacao.ufcg.edu.br) ]

Monilly Ramos Araujo Melo  [ Federal University of Campina Grande | [monillyramos@gmail.com](mailto:monillyramos@gmail.com) ]

 Departamento de Ciências Exatas, Universidade Federal da Paraíba, Avenida Santa Elisabete, s/n, Centro, Rio Tinto, PB, 58297-000, Brasil.

**Received:** 01 December 2023 • **Accepted:** 18 February 2025 • **Published:** 20 May 2025

**Abstract** Computational Thinking (CT) is a cognitive problem-solving approach commonly employed in the field of Computer Science. Over recent years, various strategies have emerged to promote CT awareness and understanding. Despite these initiatives, there has been a lack of quantitative analysis aimed at assessing CT as a cognitive skill among undergraduate students, particularly focusing on items designed for this purpose. In this study, our objective is to investigate the psychometric properties of CT questions as answered by novice Computer Science undergraduates. To achieve this, we selected a set of questions from the Bebras Challenge, an international competition designed to explore CT skills without requiring programming expertise. In pursuit of our goal, we utilized Item Response Theory (IRT) to scrutinize the difficulty and discrimination levels of these selected Bebras questions. Difficulty is related to how an examinee responds to an item, while discrimination measures how effectively an item can differentiate between individuals with higher and lower levels of knowledge. Our findings reveal several key insights: (i) Concerning the accuracy in predicting question difficulty, theoretical predictions achieved an accuracy rate ranging from 53% to 58% when compared to empirical data. (ii) The Bebras Challenge questions predominantly exhibited two levels of difficulty, spanning from easy to medium. (iii) The questions displayed a spectrum of discrimination levels, encompassing low, moderate, and high discrimination, a crucial aspect for crafting effective assessment instruments. Additionally, we have gathered observed lessons from this exploratory study, regarding the design of questions that can contribute to reliably measure CT skills. These lessons contribute to understanding features influencing the reliable design of items for measuring CT skills. These insights serve as a resource for future research endeavors aimed at enhancing our understanding of assessing CT abilities.

**Keywords:** Computational Thinking Assessment, Item Response Theory, Bebras Challenge

## 1 Introduction

Computational Thinking (CT) has garnered significant attention in the field of Computer Science Education over the past few decades. CT originated from Seymour Papert's constructionist learning ideas [Papert, 1980; Papert and Harel, 1991]. In 2006, Jannette Wing, in her seminal paper, reintroduced CT as a “fundamental, not rote skill” for problem-solving that “builds on the power and limits of computing processes” [Wing, 2006]. Wing and Papert both emphasize reasoning, albeit with distinct nuances. Wing's definition underscores problem-solving, while Papert's focuses on stimulating reflective thinking [Mannila *et al.*, 2014]. Since then, the definition of CT has evolved in the context of solving general problems through the exploration of Computer Science skills [Kalelioglu *et al.*, 2016; Ilic *et al.*, 2018; Tsarava *et al.*, 2021].

Studies, as indicated by Tang *et al.* [2020], associate CT skills with programming practices and competencies required in specific domains, as well as with general problem-solving abilities. Regarding programming practices, Brennan and Resnick [2012] propose a framework that encompasses computational concepts, computational practices, and computational perspectives. Conversely, Selby [2015] advocate for a Computational Thinking Taxonomy comprising ab-

straction, algorithm, evaluation, decomposition, and generalization [Selby, 2014]. These skills find ample applications in addressing everyday problems.

However, whether related to programming practices or not, there is a paucity of knowledge regarding how to reliably measure CT skills [Moreno-León *et al.*, 2018; Ilic *et al.*, 2018; Lockwood and Mooney, 2017]. A recent mapping study concluded that most CT assessment methods rely on self-reporting rather than robust instruments [Tikva and Tambouris, 2021]. The authors also underscore the need for validated methods. Indeed, assessing a construct or latent trait presents significant challenges [Román-González, 2015]. From a psychometric perspective, a construct is a cognitive variable that cannot be directly observed, much like time or temperature [Baker and Kim, 2017]. In our context, CT skills are constructs that cannot be assessed without a purpose-designed approach. In this regard, we reaffirm the notion that CT is a complex cognitive construct [Román-González *et al.*, 2016].

Computer programming may initially appear to be a promising avenue for understanding and measuring Computational Thinking. However, it presents several challenges [Basu *et al.*, 2020]. First and foremost, in most instances, students must acquire proficiency in computer programming

before they can write code to solve problems. The drawback of this approach lies in the fact that CT abilities are evaluated based on the programming skills of the individuals. It is essential to recognize that students are required to adopt a new means of expressing solutions, distinct from their natural language [Sherman and Martin, 2015; Giordano and Maiorana, 2014; Duncan and Bell, 2015]. Moreover, this approach can introduce bias into the construct under examination by teaching what we intend to observe.

Secondly, the assessment often involves checking whether the code produced by students contains specific structural elements from a predefined checklist [Seiter and Foreman, 2013]. Thirdly, measuring CT abilities through coding can potentially blur the lines between cognitive skills and technical skills. In other words, the challenges of learning a new programming language can impact students' ability to articulate solutions, subsequently affecting their CT abilities.

On the other hand, there is a growing field dedicated to exploring the use of non-coding instruments for assessing CT abilities. Several initiatives strive to offer guidelines for the development of tests designed to evaluate CT abilities, drawing inspiration from tasks commonly found in Bebras Challenges [Basu et al., 2020; Lockwood and Mooney, 2018; Palts et al., 2017]. Bebras<sup>1</sup> stands as an international initiative with a primary focus on promoting CT in schools. Since its inception in 2006, the Bebras community has consistently generated CT questions (tasks) designed to hone CT skills for solving general problems, all without necessitating programming or Computer Science knowledge [Dagienė and Futschek, 2008]. Notably, the Bebras Challenge operates with its proprietary system for question (task) generation and classification, categorizing them based on difficulty levels such as easy, medium, or difficult. This process involves predicting the difficulty level of questions before they are answered, drawing on stakeholder knowledge [VANÍČEK et al., 2021]. Our research endeavors to empirically examine the design of these questions and extract insights from the wealth of knowledge already produced by the Bebras community. To analyze the empirical data, we employed Item Response Theory (IRT) as a guide for delving into the dataset.

In the course of this study, our principal aim is to scrutinize the intricacies of questions designed for the evaluation of CT skills. This investigation takes a comprehensive approach by conceptualizing CT skills as a construct delineated through the prism of Bebras tasks (questions). We seek to harness the potential of tasks curated by informatics experts globally. In this context, our research endeavors to provide insights into the following pivotal research questions: *RQ1 - Accuracy in Predicting Difficulty Levels*: How effectively can we forecast the difficulty levels of Bebras tasks utilizing Item Response Theory (IRT)? *RQ2 - Utility of Bebras Tasks in CT Assessment*: To what extent do Bebras tasks serve as effective tools in assessing Computational Thinking skills, as measured by IRT metrics? *RQ3 - Best Practices for Item Development in CT Assessment*: What are the recommended practices for the development of items geared towards the assessment of Computational Thinking, particularly within

the framework of Bebras Challenge tasks? Through a meticulous exploration of these research questions, we aim to contribute nuanced perspectives and empirical evidence to the ongoing discourse surrounding the assessment of Computational Thinking skills using Bebras tasks and Item Response Theory.

Our principal contribution lies in providing a critical perspective on the design of items for assessing CT skills, focusing on two psychometric properties in Bebras framework: *difficulty* and *discrimination* [Hambleton et al., 1991]. Historically, in the Bebras Challenge, neither the *difficulty* nor the *discrimination* of tasks is a primary consideration for task designers. *Difficulty* is related to how an examinee responds to an item, while *discrimination* measures how effectively an item can differentiate between individuals with higher and lower levels of knowledge [Baker and Kim, 2017]. Both psychometric properties will be examined in this exploratory study.

The structure of this work is as follows: The *Background section* introduces CT skills, the Bebras contest, and provides a brief explanation of Item Response Theory (IRT), with a particular focus on the concepts of “difficulty” and “discrimination.” The *Related Work section* presents prior research related to assessing CT. The *Material and Method section* elaborates on the participants, instruments, procedures, and ethical principles. The section on *Evaluating Item Response Theory Properties in Tasks* presents the results and addresses RQ1 and RQ2. Subsequently, *Section 6* addresses RQ3, presents the lessons learned, and discusses potential threats to the validity of our study. Finally, we conclude the article and outline potential future research directions.

## 2 Background

This section provides the necessary background information. It introduces the concept of computational thinking (CT) skills and emphasizes the significance of the Bebras Challenge as a pivotal tool in promoting CT. Additionally, we offer an elucidation of the psychometric properties utilized in our methodology, delineating the fundamental concepts of Item Response Theory.

### 2.1 Computational Thinking Skills

Numerous studies have been dedicated to identifying the foundational skills that define computational thinking (CT). According to Wing [2008], abstraction is the crux of CT, as engaging with multiple layers of abstraction aids in comprehending both problems and solutions. Barr and Stephenson [2011] outline a comprehensive list of core CT concepts and capabilities for K-12 education, encompassing data collection, analysis, and representation, problem decomposition, abstraction, algorithms and procedures, automation, parallelization, and simulation. Grover and Pea [2013] propose an extensive range of capabilities for CT, including abstractions and pattern generalizations, systematic information processing, symbol systems and representations, algorithmic flow control, structured problem decomposition, iterative, recursive, and parallel thinking, conditional logic, efficiency con-

<sup>1</sup>Bebras International Challenge on Informatics and Computational Thinking. <https://www.bebas.org/>

siderations, and systematic error detection. These diverse abilities underscore the fluid nature of CT skills, which lack distinct demarcations.

The International Society for Technology in Education (ISTE) and the Computer Science Teachers Association (CSTA) collaborated on establishing an operational definition for K-12 education. This definition has become instrumental as a guiding framework for various studies in the field. According to these educational associations, CT is a problem-solving process characterized by the following attributes ISTE [2011]:

- “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;”
- “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.”

They also endorse the CT operational definition with a set of attitudes that includes ISTE [2011] (page 1): (i) “Confidence in dealing with complexity; (ii) Persistence in working with difficult problems; (iii) Tolerance for ambiguity; (iv) The ability to deal with open ended problems; and (v) The ability to communicate and work with others to achieve a common goal or solution”.

Related to the competencies necessary for general problem-solving skills, the Computer at School project defines *algorithm thinking*, *decomposition*, *generalization*, *abstraction*, and *evaluation* as fundamental concepts of computational thinking [Csizmadia et al., 2015]. *Algorithm thinking* entails the creation of ordered steps aimed at achieving a particular objective, encompassing the adherence to sequences and rules that lead to a defined goal. The cognitive skills associated with algorithmic thinking include understanding, devising, and implementing [Dagienė et al., 2017].

The concept of *decomposition* revolves around the ability to approach a problem by analyzing its constituent parts. Through the process of problem decomposition, the overarching goal is dissected into more manageable components, thereby facilitating easier resolution.

*Generalization*, on the other hand, refers to the capacity to identify recurring patterns within problem-solving contexts. For instance, individuals can leverage their past problem-solving experiences to adapt to similar situations. This capacity often involves inductive reasoning, wherein a generalized solution is applied to achieve a specific goal [Dagienė et al., 2017].

In the context of problem-solving, *abstraction* pertains to the identification of essential elements necessary to resolve a given issue by eliminating extraneous details. The primary

objective of abstraction is to simplify the complexity associated with a problem [Barr and Stephenson, 2011].

Lastly, *evaluation* encompasses the ability to determine the optimal solution while considering various constraints such as time, distance, or available resources. It also involves reflecting on the means of achieving or representing the solution [Dagienė et al., 2017].

## 2.2 Bebras Challenge

Bebras represents an international initiative that operates under the premise that the teaching of computer programming is not obligatory to foster computational thinking [Dagienė et al., 2016]. At its core, the Bebras challenge encompasses an annual competition, conducted one or two times a year across six distinct age groups. While the contest is frequently administered in schools via computers, it can also be undertaken using traditional pen-and-paper methods. During the competition, the participants are supervised by teachers, who may seamlessly integrate the event into their pedagogical activities [Dagienė et al., 2016]. Each pupil is required to solve between 15 to 18 tasks within a time frame of 40 to 55 minutes.

An integral aspect of the Bebras Challenge pertains to the orchestration of the contest. Annually, researchers from participating countries convene in workshops to devise and refine the tasks for the competition. These tasks are designed to be succinct, answerable within a few minutes via a digital interface, and necessitate profound cognitive engagement within the field of informatics [Dagienė et al., 2015]. The tasks themselves are either multiple-choice, featuring four options with one correct answer, or interactive, involving techniques such as drag-and-drop, construction assembly, item selection, or text input [Dagienė et al., 2014]. As an illustration, Figure 1 exemplifies a Bebras task named “Space Maze”. Notably, the tasks are initially crafted in the English language, and it falls upon each national organizer to translate the tasks into their respective native languages.

Despite the fact that Bebras challenges can be tackled without prior knowledge of Computer Science, all questions are indirectly linked to fundamental computational concepts. Over the years, the categories of contents have undergone modifications, resulting in a contemporary two-dimensional classification system [Dagienė et al., 2017]. The first classification encompasses the categories of *Informatics Concepts* at a knowledge level, comprising five core topics, namely: (i) Algorithms and programming, including logical reasoning; (ii) Data, data structures, and representations, encompassing graphs, automation, and data mining; (iii) Computer processes and hardware, encompassing various aspects of computer functionality such as scheduling and parallel processing; (iv) Communications and networking, involving subjects like cryptography and cloud computing; and (v) Interaction (Human-Computer Interaction, HCI), systems, and society, encapsulating miscellaneous topics [VANÍČEK et al., 2021]. The second classification encompasses the categories of Computational Thinking at a skills level, including five key abilities, namely: (i) Abstraction, (ii) Algorithmic thinking, (iii) Decomposition, (iv) Evaluation, and (v) Generalization.

Another notable category pertaining to the tasks is their level of difficulty. The tasks are organized into three distinct levels, representing increasing degrees of complexity (easy, medium, and hard). These classifications are determined by the stakeholders based on their preliminary intuition prior to the commencement of the contest [VANÍČEK *et al.*, 2021; Van der Vegt, 2018]. These difficulty levels determine the score assigned to a task when answered correctly. For tasks classified as easy, the score awarded for a correct answer is 6 points. For medium-difficulty tasks, the score is 9 points, and for difficult tasks, it is 12 points. The final test score is the sum of these points.

As an example, Figure 1, the “Space Maze” task, was classified as an easy question by the United Kingdom organizers, who additionally highlighted algorithmic thinking as an essential skill required to tackle the task. The answer to the question in Figure 1 is option A.

Some space explorers landed on an empty planet. From their ship they could see a maze with an unknown golden object in it. The explorers dropped their robot into the maze hoping it could take a closer look at the unknown object. Unfortunately the robot broke during the fall and can now only send and receive garbled instruction about where to go.

The robot suggests four possible directions it can go. Even though the words in the instructions are garbled, there are still only four different words, each indicating north, west, east or south. When following the instructions the robot will move into an adjacent square as instructed.

**Which instructions should the explorers send the robot in order for it to reach the golden object?**

A. Ha' poS poS Ha' Ha' nIH  
 B. Ha' poS poS Ha' nIH Ha'  
 C. Ha' Ha' poS Ha'  
 D. Ha' poS nIH vI'ogh Ha' poS




Figure 1. Space Maze task

## 2.3 Item Response Theory

Psychometric theories have found application in the assessment of latent variables [Erthal, 1987]. Although a latent variable cannot be directly observed, its existence can be inferred from observable or manifest variables. Item Response Theory (IRT) serves as a method for devising, analyzing, and scoring instruments aimed at quantifying latent variables, such as abilities or attitudes [Hambleton *et al.*, 1991].

IRT, as a mathematical testing model, centers on an individual's performance on a test designed to measure specific abilities [Baker, 2001]. Baker elucidates that “the foundational principles of item response theory are rooted in the individual items of a test, rather than in an aggregation of item responses like a test score” [Baker, 2001] (page 6). In other words, whereas Classical Test Theory (CTT) is predicated on the total score, IRT operates on the premise that an examinee's response to each item, rather than the overall score, serves as the baseline for analysis.

According to IRT, an instrument can be evaluated based on three key criteria [Baker, 2001]. The first criterion is the **discrimination** of an item, which assesses the item's ability to differentiate between individuals with varying levels of knowledge. An item with higher discrimination demonstrates an enhanced capacity to discern even slight changes in an examinee's ability level. The second criterion is the **difficulty** of the item, which situates the item along the ability scale. For instance, an easy item is tailored for individuals with lower abilities, whereas a difficult item is designed for those with higher abilities. Finally, the third parameter is referred to as the “guessing parameter.” This parameter accounts for the probability that an individual may answer a question correctly purely by guessing without having actual knowledge or mastery of the subject. In the context of multiple-choice or short-answer tests, this phenomenon occurs when an individual selects or suggests a correct answer based on intuition or random choice rather than understanding. The guessing parameter is particularly relevant for multiple-choice questions, where the likelihood of guessing correctly increases with fewer response options.

The three-parameter logistic model (3PL) is used for analyzing difficulty, discrimination, and probability of hit due to guessing. Equation 1 shows the fundamental equation of the 3PL model [Baker, 2001]. The Greek letter  $\Theta$  represents the examinee ability in IRT. Equation 1 describes the probability that an examinee with  $\Theta$  will correctly answer an item  $j$ , with a discrimination level  $a$  (the slope parameter), a difficulty level  $b$  (the threshold parameter), and the probability of guessing  $c$  (lower asymptote parameter) [Baker, 2001]. In the equation,  $e$  is the base of the natural logarithm constant approximately equal to 2.718. The theoretical range of ability  $\Theta$  scale is infinity, but frequently, the practical range is from -3 to +3 [De Ayala, 2013].

$$P(\Theta) = c_j + (1 - c_j) \frac{1}{1 + e^{-a_j(\Theta - b_j)}} \quad (1)$$

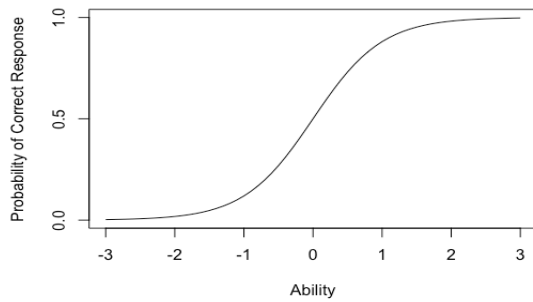
There are the two-parameter logistic model (2PL) and the one-parameter logistic model (1PL). The 2PL represents the equation of item parameters  $b$  (difficulty) and  $a$  (discrimination), as shown in Equation 2 [Baker, 2001]. Therefore, 2PL defines the discrimination and difficulty level of item, excluding the guessing ( $c$  parameter).

$$P(\Theta) = \frac{1}{1 + e^{-a_j(\Theta - b_j)}} \quad (2)$$

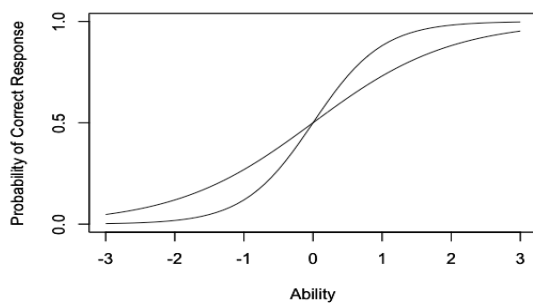
Finally, the one-parameter logistic model (1PL) or Rasch Model characterizes only the item difficulty level. Under the Rasch Model, the discrimination parameter is fixed at a value of  $a = 1$  for all items. Thus, only the difficulty parameter  $b$  is estimated, as shown in Equation 3. As we can notice, the equations of 3PL, 2PL, and Rasch model are similar, but it depends on which parameters we want to estimate.

$$P(\Theta) = \frac{1}{1 + e^{-1(\Theta - b_j)}} \quad (3)$$

Another crucial component in the framework of Item Response Theory (IRT) is the Item Characteristic Curve (ICC), serving as a graphical representation of difficulty and discrimination parameters. The S-shaped curve visually eluci-



**Figure 2.** Typical Item Characteristic Curves



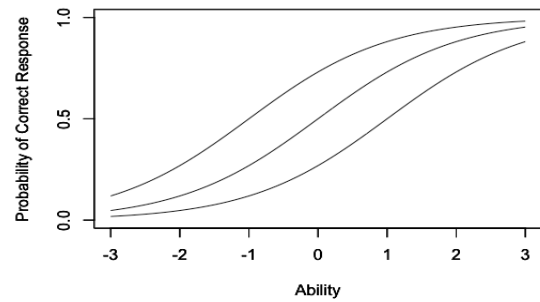
**Figure 3.** Two ICC with the same difficulty level but different discrimination

dates the correlation between an individual's ability (x-axis) and the probability of selecting the correct answer (y-axis) for a particular item [Baker, 2001]. Figure 2 depicts a typical ICC, with the difficulty level ( $b = 0$ ) and discrimination ( $a = 1$ ) parameters specified. The discrimination parameter ( $a$ ) indicates the steepness of the item, signifying how effectively the item discriminates between individuals of varying abilities in terms of the likelihood of responding correctly.

Paying close attention to the behavior of ICCs is of paramount importance. Figure 3 illustrates two ICCs with the same difficulty level but differing levels of discrimination. When the discrimination level exceeds a moderate threshold, the item characteristic curve exhibits an S-shaped pattern with a pronounced slope in its central segment. Conversely, when the item discrimination is below the moderate threshold, the ICC assumes an almost linear trajectory, appearing relatively flat.

Both Figures 3 and 4 showcase the measurement of the difficulty parameter on the same ability scale  $\Theta$  at the point where the probability of selecting the correct answer is 0.5 (or 50%) [Baker, 2001]. In both cases,  $b = 0$ . Figure 4 exemplifies three ICCs, each possessing the same discrimination parameter but varying difficulty levels. When an item is categorized as easy, the difficulty parameter is associated with a lower level of ability (leftmost curve), whereas a hard item corresponds to a higher level of ability (rightmost curve).

The discrimination parameter value  $a$  can be classified according to an item's power to differentiate examinees with higher and lower ability levels. Usually, the probability of endorsing the correct answer increases as the ability level increases. Table 1 shows the range of values for seven labels based on Baker and Kim [2017]. For negative discriminate



**Figure 4.** Three ICC with the same discrimination but different difficulty level

values, the probability of endorsing the correct answer decreases as the ability level increases.

**Table 1.** Labels for item discrimination parameter values [Baker and Kim, 2017]

Label	Range of values	Typical value
None	0	0.00
Very low	.01 - .34	0.18
Low	.35 - .64	0.50
Moderate	.65 - 1.34	1.00
High	1.35 - 1.69	1.50
Very high	> 1.70	2.00
Perfect	+ infinity	+ infinity

When examining the values of discrimination, difficulty, and item guessing parameters, it is imperative to be attentive to critical thresholds. Several issues may arise: (i) when the discrimination parameter falls below 0.30, (ii) when the difficulty parameter is less than  $-2.95$  or exceeds  $2.95$ , and (iii) when the probability of a correct response due to guessing surpasses 0.35 [Baker and Kim, 2017; Vendramini and Dias, 2005]. In such instances, it is prudent to subject the item to scrutiny, and it may be necessary to consider its removal from the test.

While there is not a universally established classification system for item difficulty, we can adopt the interpretations offered by Hambleton *et al.* [1991] and Baker and Kim [2017]. Given that the item difficulty represents a location parameter along the ability scale, Hambleton *et al.* [1991] assert that an item with  $b < -2$  is considered very easy, while an item with  $b > +2$  is categorized as very hard. Baker and Kim [2017] suggests that an item with a difficulty of  $-1$  is suited for individuals with lower abilities, whereas an item with a difficulty of  $+1$  is more appropriate for those with higher abilities. Consequently, it can be deduced that the range between  $b > -1$  and  $b > +1$  encompasses items of moderate difficulty. Furthermore, difficulty levels between  $b > -2$  and  $b < -1$  are indicative of easy items, whereas those between  $b < +1$  and  $b > +2$  signify hard items.

### 3 Related Work

Several studies have extensively explored Bebras tasks over the years. For instance, Van der Vegt [2013] examines the difficulty of Bebras tasks through the lens of success rates,

while employing elements of the cognitive load theory to analyze the complexity of tasks in the Netherlands [Van der Vegt, 2018]. Yagunova *et al.* [2015] delve into the intricacies of Russian contests, considering factors such as the number of correctly solved tasks and the associated cognitive, informational, and emotional loads. Pohl and Hein [2015] propose recommendations for enhancing the quality of tasks, advocating for succinct sentences, precise definitions, and unambiguous language.

Moreover, various studies have employed IRT to analyze Bebras tasks, highlighting their potential in evaluating CT skills. Lonati *et al.* [2017] not only examine Italian contest tasks under the purview of IRT but also compare the outcomes with data from modified tasks. Bellettini *et al.* [2015] explore the prediction of difficulty levels by organizers in Italy, subsequently investigating the actual difficulty using IRT analysis post-contest. Dolgopolas *et al.* [2015] utilize Bebras tasks to measure CT skills among novice software engineering students, emphasizing the significance of applying IRT for such assessments. Hubwieser and Mühling [2015] develop a specialized approach involving distinct mathematical models of IRT to scrutinize tasks within the German Bebras Challenge.

In addition, we find pertinent discussions on instruments employed for assessing CT skills, notably those that employ the Bebras challenge as a measure. Mooney and Lockwood [2020] devise two CT tests for undergraduate students in Ireland, noting no substantial differences between pre- and post-programming class assessments. Matsuzawa *et al.* [2018] utilize Bebras as a measurement tool for CT skills in an introductory programming course designed for non-computer science undergraduates, correlating the results with practical and written programming tests.

However, challenges arise in correlating Bebras tests with academic performance in structured programming courses, as evidenced by studies conducted by Dolgopolas *et al.* [2015] and Boom *et al.* [2018]. While Dolgopolas *et al.* [2015] find no significant correlation between Bebras tests and structured programming grades, Boom *et al.* [2018] establish a substantial correlation between Bebras and an intelligence test, highlighting its efficacy in measuring general problem-solving skills.

Furthermore, two validated tests have garnered attention for assessing CT skills, each focusing on different aspects of computational thinking. The Computational Thinking Test (CTt) addresses fundamental computational concepts and is administered online to Spanish students aged twelve to fourteen [Román-González, 2015]. The CTt has recently been adapted for ten-year-old primary school children [Tsarava *et al.*, 2021]. On the other hand, the CT Scale evaluates five CT factors, including algorithmic thinking, critical thinking, creativity, cooperativeness, and problem-solving, through a five-point Likert scale administered to graduate and undergraduate students [Korkmaz *et al.*, 2017]. These instruments offer comprehensive assessments of CT skills independent of specific programming environments, with the CTt focusing on programming and algorithmic problem-solving skills, while the CT Scale emphasizes broader cognitive abilities.

## 4 Material and Methods

In this section, we outline the instruments used, describe the participants involved, and delineate the procedures adopted for our study. Finally, we present our ethical principles.

### 4.1 Instruments

Our research employed Bebras tasks as the primary investigative instrument, a choice made for several compelling reasons. Firstly, these tasks are specifically designed to assess CT skills, eliminating the need for prior knowledge in Computer Science or programming. This characteristic enables students to respond to the tasks without exposure to formal Computer Science classes. Additionally, the tasks are characterized by their independence, self-contained nature, and self-explanatory structure.

Secondly, the Bebras tasks undergo a meticulous development process, initially designed by researchers worldwide and further refined during an annual workshop attended by specialists. This iterative improvement process involves critical analysis by international professionals, contributing to the production of high-quality tasks.

Finally, the Bebras Challenge organizers in various countries compile brochures for their national contests, encompassing the tasks along with their corresponding solutions. This approach facilitates the utilization of these tasks by researchers worldwide for diverse research objectives.

From the array of available brochures, we specifically chose the national Bebras Challenge from the United Kingdom (UK) in 2015 and 2014. The decision to select a foreign country's brochures stems from the fact that, up until the completion of this research in 2020, Brazil had not formally engaged in the competition. Consequently, there were no questions available in Brazilian Portuguese.

Given that Bebras tasks cater to students of various age groups, we specifically selected tasks designed for the older age group (over 16 years old) to align with the participants' age in our study. Instrument One comprises tasks from the 2015 contest, while Instrument Two comprises tasks from the 2014 contest. The task names are detailed in Table 2 and Table 3. Subsequent to the selection of instruments, a meticulous translation of the tasks into Portuguese was undertaken. The translation process consisted of the following steps: First, the tasks were translated from English to Portuguese, emphasizing incorporating linguistic and cultural adaptations where appropriate. Next, each task was reviewed by a minimum of two professors and researchers specializing in Education and Computer Science. In cases where discrepancies arose in the translation, a third professor was consulted for clarification. Overall, six volunteer professors and researchers evaluated at least three tasks each. The tasks are available in the Appendix.

### 4.2 Participants

Our study comprised 214 novice undergraduate students, aged 17 to 24, who were enrolled in an introductory programming course at two Federal Universities in Brazil during the academic years 2017 and 2018. These participants



**Table 2.** Tasks of instrument One

Item	tasks	Item	tasks
item 01	Drawing stars	item 07	Word chain
item 02	Bowl Factory	item 08	Fireworks
item 03	Email	item 09	You won't find it
item 04	Beaver the alchemist	item 10	Turn the cards
item 05	Tutorial	item 11	Decorating chocolate
item 06	Popularity	item 12	Busy beaver

**Table 3.** Tasks of instrument Two

Item	Tasks	Item	Tasks
item 01	Ceremony	item 09	Social network
item 02	Log-art	item 10	Height game
item 03	Beavers on the run	item 11	Meeting point
item 04	Traffic in the city	item 12	Best translation
item 05	Storm proof network	item 13	Broken machines
item 06	Space maze	item 14	True or false
item 07	Footprints	item 15	Right rectangles
item 08	Puddle jumping		

were novices in Computer Science, specifically in the realm of programming. They participated in our study before before starting formal programming classes at the university. The introductory programming courses were structured with the primary objective of fostering the development of students' programming skills, with a specific emphasis on utilizing Python as the programming language.

While Bebras primarily caters to pupils in schools, its tasks have been employed in research involving higher education students [Mooney and Lockwood, 2020]. This occurs when selecting Bebras questions designed for older audiences, specifically those aged 16 and above, as is the case in our study.

Prior to the commencement of our study, all students or their legal representatives actively participated in the research by signing informed consent forms. This ethical measure ensured that the participants were fully aware of and consented to their involvement in the study. To protect the confidentiality and privacy of the participants, all collected data were anonymized. This deliberate step was taken to safeguard the identities of the students, making it impossible to discern or identify individual examinees from the dataset.

### 4.3 Procedures of data collection and analysis

The instruments were distributed in a printed format, and students responded to the questions using traditional pen-and-paper methods. The primary researcher oversaw the data collection process, where, in each instance, the objectives of the research and ethical considerations were thoroughly explained to the participants. The time allocated for the completion of the tests, set at 55 minutes, adhered to the guidelines outlined in the Bebras Challenge instructions.

Following the data collection phase, responses were meticulously organized into a table. Each correct answer received a score of one (+1), while incorrect answers were assigned a score of zero (0) in both instruments. Subsequently, the total score for each student was calculated based on the aggregate number of correct responses. Notably, we diverged from the point counting system employed by the Bebras Challenge, as our analysis focused on dichotomous items—responses were categorized strictly as correct or incorrect.

Given our primary objective of scrutinizing how items

were designed to assess computational thinking (CT) abilities, we employed Item Response Theory (IRT) for a comprehensive evaluation. Emphasizing the advantages of IRT, we specifically concentrated on two pivotal criteria: the difficulty and discrimination parameters of the items. These elements play a crucial role in elucidating how effectively an item measures cognitive abilities in examinees [Baker, 2001; Baker and Kim, 2017; Hutz *et al.*, 2015; Pasquali, 2017].

Our subsequent analysis involved a detailed examination of difficulty and discrimination using the IRT methodology. We utilized the add-on Eirt for Microsoft Excel<sup>2</sup>. Eirt, developed with the Item Response Theory Library (libirt) in C language under the GNU Public License, facilitates the estimation of item parameters and examinee abilities using one, two, or three logistic models (Rasch, 2PL, or 3PL). The parametric estimator employed is the Bayes Modal Estimator, while the ability estimator is the Expected a Posteriori (EAP).

Our first research question (RQ1) delves into the accuracy of predicting difficulty levels. Unlike the Bebras Challenge, where difficulty level classification relies on subjective judgments and tacit knowledge of the organizers, we advocate for a data-driven approach. Therefore, we perform a difficulty level classification based on the difficulty parameter under IRT, contrasting the classifications made by organizers before and after students answered the tasks. The accuracy of predicting difficulty levels is then calculated to address RQ1.

Moving on to our second research question (RQ2), we explore the utility of tasks in assessing CT by scrutinizing their psychometric properties under IRT, namely the difficulty and discrimination parameters. Notably, in the Bebras Challenge, these parameters are not the primary focus of task designers. Therefore, we adopt them as criteria to evaluate the suitability of tasks in assessing CT skills.

Our third research question (RQ3) investigates best practices for producing items to assess CT. We address this question by presenting a critical perspective on item design based on observations during the IRT analysis. As tasks were scrutinized, we systematically gathered and aggregated their characteristics, considering both discrimination and difficulty parameters. The result is a compilation of lessons learned, providing valuable insights into the process of item design for assessing CT.

### 4.4 Ethical Principles

The research project received approval from the Ethics Committee of the Federal University of Campina Grande. The study was officially registered on Plataforma Brasil under the identification number CAAE: 56847316.7.0000.5182. All study participants or their legal representatives provided signed informed consent forms prior to the beginning of the study. To uphold the confidentiality of the subjects, the data were anonymized, making it impossible to identify the examinees.

<sup>2</sup>Eirt - Available at <http://psychometricon.net/libirt/>

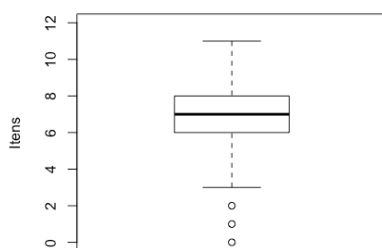
## 5 Evaluating Item Response Theory Properties in Tasks

This section presents the outcomes pertaining to the difficulty and discrimination levels of items as analyzed through Item Response Theory (IRT). We provide visual representations, statistical findings, and a comprehensive classification of both difficulty and discrimination, all grounded in the principles of IRT. Finally, the synthesis of these results contributes to addressing Research Questions 1 (RQ1) and 2 (RQ2).

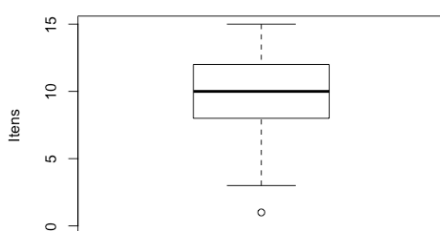
### 5.1 Data Screening and Statistical Procedures

With the exception of a few outliers, the age range of all participants during the data collection period fell between 17 and 21 years. For Instrument One data collection, the sample comprised 170 students, with 150 males and 20 females. Figure 5a illustrates the boxplot of total scores, ranging from 0 to 12 correct answers. The mean score was 7.08, with a standard deviation of 2.04. In the context of Instrument Two data collection, there were 214 participants, consisting of 188 males and 26 females. The distribution of data is depicted in the boxplot shown in Figure 5b, where the total score spans from 0 to 15 correct answers. The mean score was 9.701, with a standard deviation of 2.63.

It is noteworthy that the same cohort of students who participated in Instrument One also responded to Instrument Two. Additionally, a subset of 44 students who were absent during Instrument One's data collection due to unforeseen circumstances subsequently participated in Instrument Two. Consequently, the total number of students involved in our research amounted to 214. Importantly, the varying number of participants between the two instruments is inconsequential to the analysis, as we conducted separate analyses for each instrument, ensuring the robustness and relevance of our results.



(a) Boxplot of total scores from Instrument One



(b) Boxplot of total score from Instrument Two

**Figure 5.** Boxplots of Instruments

procedures to verify that the data collected from students at both universities could be combined into a single database. In other words, it was important to determine whether the data from distinct groups, such as students from different universities, came from the same distribution. This step was essential to facilitate the integration of students' responses from two universities into a cohesive dataset that encompasses data from both Instrument One and Instrument Two. To achieve this, we performed multiple hypothesis tests using various combinations of students' responses from different universities for each instrument. The Chi-squared two-sample test was employed, with a significance level set at 0.05. The null hypothesis stated that the two samples emanated from a common distribution, while the alternative hypothesis suggested otherwise. The chosen significance level of 0.05 remained consistent across all tests. The results of the Chi-squared two-sample test yielded a  $p$ -value greater than 0.05 for all sample combinations. Consequently, we embraced the null hypothesis, indicating that the samples originated from a common distribution ( $p$ -value  $> 0.05$ ). This confirmation allowed the data to be organized into a common database.

Before presenting the results, we initiated the statistical procedures by selecting the appropriate logistic model (Rasch, 2PL, or 3PL). The Rasch (or 1PL) model considers only the difficulty level, the 2PL incorporates both discrimination and difficulty levels, and the 3PL model additionally considers correct answers by guessing, beyond the previously mentioned parameters. A criterion for model selection is to opt for the one with lower values of AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion), as outlined in the protocol described by Baker and Kim [2017]. Table 4 presents the AIC and BIC values for Rasch and 2PL models, as well as the 2PL and 3PL models for Instrument One. Similarly, Table 5 provides the corresponding values for Instrument Two. Additionally, we conducted an analysis of variance (ANOVA) to assess whether the differences between these models are statistically significant.

**Table 4.** Fit of models - Instrument One

Model	AIC	BIC	Model	AIC	BIC
Rasch	1724.207	1763.171	2PL	1717.394	1789.327
2PL	1717.394	1739.327	3PL	1721.677	1829.577

**Table 5.** Fit of models - Instrument Two

Model	AIC	BIC	Model	AIC	BIC
Rasch	2717.754	2768.299	2PL	2715.977	2810.749
2PL	2715.977	2710.749	3PL	2724.253	2866.411

To summarize, the determination of acceptable models hinges on the lower values of AIC and BIC. Across both Instruments, the 2PL model emerged as the most fitting choice. As depicted in Tables 4 and 5, both AIC and BIC consistently exhibited lower values for the 2PL model. Furthermore, the ANOVA results revealed a statistically significant difference between the Rasch and 2PL models ( $p$ -value = 0.002 for Instrument One and  $p$ -value = 0.008 for Instrument Two). However, no statistically significant difference was found between the 2PL and 3PL models ( $p$ -value = 0.073 for Instrument One and  $p$ -value = 0.115 for Instrument Two). It is important to note that the significance level was

Before starting the IRT analysis, we performed statistical



set at 0.05 in all cases. Consequently, we proceeded with the IRT analysis employing the 2PL model for both Instruments.

## 5.2 IRT Results

Our exploration of IRT parameters commenced with an examination of the Item Characteristic Curve (ICC) for Instrument One, as illustrated in Figure 6. Throughout the subsequent discussion, we refer to *tasks* as *items* for consistency across both instruments.

Our observations unveiled distinct patterns in the ICC. Notably, Item 3, denoted as the “Email” task, emerged not only as the easiest item, evident from its position as the leftmost curve, but also exhibited one of the lowest discrimination values, indicated by the minimal slope of its ICC. Remarkably, the probability of providing correct answers extended up to 50%, even among individuals with lower ability levels.

Conversely, Item 10, referred to as “Turn the cards,” represented the most challenging item, displaying the lowest discrimination value characterized by an almost linear curve. A similar curve behavior was noted in Item 8, the “Fireworks” task, where the ICC exhibited a nearly linear trajectory. Consequently, Items 3, 8, and 10 were identified as having the lowest discrimination values among all tasks.

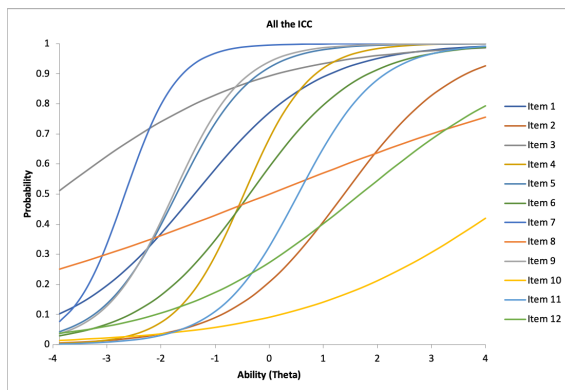


Figure 6. Item Curve Characteristics of Instrument One

Table 6 shows the calculated difficulty and discrimination parameters of Instrument One. In Table 6, our visual assessment, as discussed earlier, is substantiated through a closer examination of the difficulty and discrimination parameters. A meticulous scrutiny of the difficulty parameter (column *b*) reveals two items with noteworthy concerns, identifiable by critical values wherein the threshold value *b* falls below  $-2.95$  or exceeds  $2.95$ .

Specifically, Item 3, designated as the “Email tasks,” emerges as notably facile ( $b = -3.964$ ), aligning with its visual representation on the ICC situated further to the left in Figure 6. Conversely, Item 10, titled “Turn the cards,” poses a significant challenge ( $b = 4.656$ ), consistent with its corresponding ICC positioned more to the right in the figure. While the Bebras organizers accurately classified the “Email” task as easy, their assessment underestimated the difficulty of the “Turn the cards” task.

It is crucial to note that these findings do not necessarily indicate effective questions for assessing computational thinking skills. The “Turn the cards” task, for instance, involves logic reasoning implication, and the complexity of

this implication may have been misunderstood. Similarly, the “Email task” revolves around an apparent internet fraud scenario aimed at exploiting naive individuals for monetary gain.

Table 6. Parameters of Instrument One

Item	Tasks	a	b
1	Drawing stars	0.873	-1.38
2	Bowl Factory	0.967	1.384
3	Email	0.533	-3.964
4	Beaver the alchemist	1.644	-0.466
5	Tutorial	1.435	-1.712
6	Popularity	0.995	-0.361
7	Word chain	2.064	-2.664
8	Fireworks	0.282	0.015
9	You won't find it	1.558	-1.767
10	Turn the cards	0.494	4.656
11	Decorating chocolate	1.359	0.539
12	Busy beaver	0.581	1.689

Transitioning to the discrimination analysis, it is apparent that Instrument One exhibits a well-distributed range of discrimination levels. The classification, as per Baker and Kim [2017], is detailed in Table 7. Notably, the spectrum of discrimination values spans from very low to very high for our dataset.

Upon closer examination, one item stands out with a particularly low discrimination value, while another item registers a remarkably high discrimination value. Following this trend, three items demonstrate lower discrimination, whereas the remaining three exhibit a moderate level of discrimination. Finally, four items showcase high discrimination values. This balanced distribution across various discrimination levels contributes to the robustness and diversity of the instrument’s evaluative capacity.

Table 7. Discrimination of Item by IRT - Instrument One

Item	Tasks	Discrimination
1	Drawing stars	Moderate
2	Bowl Factory	Moderate
3	Email	Low
4	Beaver the alchemist	High
5	Tutorial	High
6	Popularity	Moderate
7	Word chain	Very High
8	Fireworks	Very Low
9	You won't find it	High
10	Turn the cards	Low
11	Decorating chocolate	High
12	Busy beaver	Low

In addition to assessing discrimination levels, we can gain insights into the precision of ability estimation by examining the information function, as illustrated in Figure 7 [Baker and Kim, 2017]. Notably, Item 7 emerges with the highest information content, particularly at an ability level of approximately  $-2.5$ . This finding indicates that the “Word chain” task is particularly adept at distinguishing individuals with very low abilities.

Contrastingly, Item 4 exhibits maximum information at an ability level close to zero, spanning from  $-1.5$  to  $+1.5$ . This implies that the “Graph coloring” task effectively discerns individuals with abilities across this mid-range spectrum. Items 9 and 5 present a comparable amount of infor-

mation for individuals with low abilities, while Item 11 offers information across a broader range of medium to high abilities ( $-1 < \theta < +2$ ).

It is worth highlighting that, among all the questions, both Item 10 and Item 3 yield a lower amount of information. Specifically, Item 10 is less informative in high ability levels, while Item 3 exhibits lower information content in low ability levels. These nuances in the information function shed light on the differential precision of each task in estimating abilities across distinct proficiency levels.

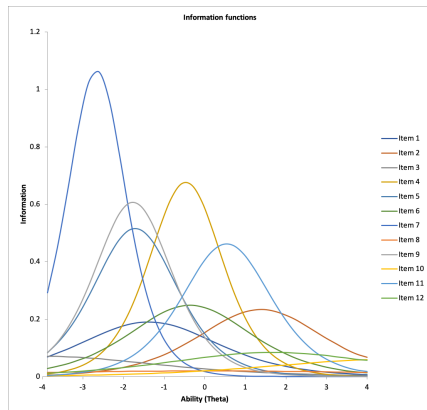


Figure 7. Information function of Instrument One

Figure 8 offers an initial insight into the Item Characteristic Curves (ICCs) as we commence the analysis of Instrument Two. Notably, Item 1, positioned on the far right and titled “Ceremony” task, emerges as the most challenging question. In contrast, Item 2, labeled “Log-art,” and Item 3, denoted as “Beavers on the run,” are identified as the easiest questions.

Items 1 and 2 exhibit nearly straight curves, indicative of low discrimination values. Specifically, Item 3 presents a distinctive characteristic with a probability of almost 30% for correct responses even from individuals with low abilities. This nuanced analysis of the ICCs offers valuable insights into the relative difficulty and discrimination levels of each item in Instrument Two.

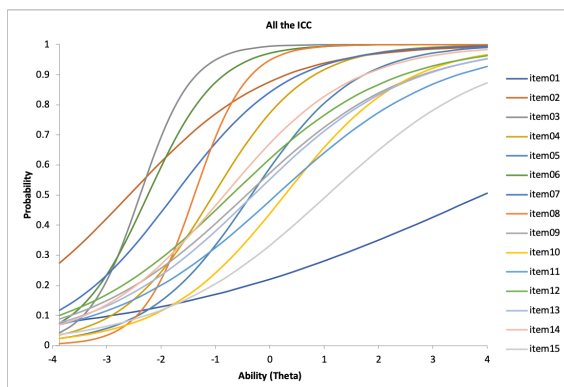


Figure 8. Item Curve Characteristics of Instrument Two

The parameters for Instrument Two are presented in Table 8. Upon scrutinizing the difficulty values of the items, a critical observation surfaces with respect to Item 1, known as the “Ceremony” task, where the difficulty parameter ( $b = 3.921$ ) exceeds the typical range (between  $-3$  and  $+3$ ). This finding implies that the question may warrant exclusion from the instrument due to its anomalous difficulty level. Additionally,

although Item 2 does not exhibit critical values, the nearly linear shape of its Item Characteristic Curve (ICC) in Figure 8 suggests that a more thorough review of this item may be warranted.

Table 8. Parameters of Instrument Two

Item	Task	a	b
1	Ceremony	0.321	3.921
2	Log-art	0.759	-2.593
3	Beavers on the run	2.103	-2.389
4	Traffic in the city	1.172	-1.042
5	Storm proof network	1.052	-0.345
6	Space maze	1.556	-2.244
7	Footprints	0.95	-1.762
8	Puddle jumping	2.081	-1.39
9	Social network	0.676	-0.437
10	Height game	0.899	0.275
11	Meeting point	0.653	0.118
12	Best translation	0.692	-0.707
13	Broken machines	0.697	-0.304
14	True or false	0.852	-0.839
15	Right rectangles	0.655	1.056

Upon examining the discrimination values, it is evident that Instrument Two comprises questions with predominantly moderate discrimination values. The classification of items based on discrimination values is elucidated in Table 9. Notably, a majority of the items, specifically eleven out of thirteen (73%), demonstrate moderate discrimination. Additionally, one item registers a very low discrimination value (Ceremony task), another exhibits high discrimination (Space Maze task), and two items showcase very high discrimination values (Puddle Jumping and Beavers on the Run tasks).

Given that the item with very low discrimination is also the one recommended for exclusion due to critical difficulty values, the overall instrument showcases items with moderate to high discrimination values. Considering the medium difficulty level classification assigned to the instrument, we can infer that it is adept at distinguishing individuals with moderate levels of ability (Theta).

Table 9. Discrimination of item by IRT - Instrument Two

Item	Tasks	Discrimination
1	Ceremony	Very low
2	Log-art	Moderate
3	Beavers on the run	Very high
4	Traffic in the city	Moderate
5	Storm proof network	Moderate
6	Space maze	High
7	Footprints	Moderate
8	Puddle jumping	Very high
9	Social network	Moderate
10	Height game	Moderate
11	Meeting point	Moderate
12	Best translation	Moderate
14	True or false	Moderate
15	Right rectangles	Moderate

Upon inspecting the information functions depicted in Figure 9, we can discern the distribution of information across various ability levels for each item. A comprehensive overview reveals that Items 3, 8, 6, and 4 contribute significant information at the lower end of the ability spectrum. In contrast, Items 5, 10, 12, and 13 yield substantial information within the medium ability range.

It is crucial to underscore that Item 1, denoted as the “Ceremony” task, stands out with the lowest amount of information, approaching zero. This signifies that Item 1 may not be well-suited for effective assessment purposes due to its lim-

ited capacity to differentiate between individuals with varying abilities.

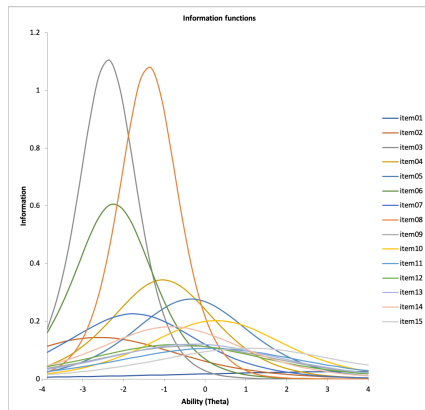


Figure 9. Information function of Instrument Two

### 5.3 (RQ1) - Accuracy in Predicting Difficulty Levels: How effectively can we forecast the difficulty levels of Bebras tasks utilizing IRT?

Table 10 delineates the difficulty level classification for Instrument One according to both IRT and Bebras' organizers. A close examination of the Bebras organizers' difficulty predictions in comparison to the IRT parameters reveals a precision rate of only 58%. The classification entails five tasks labeled as easy (with two of them falling under the category of very easy), four as medium, and three as hard (including one categorized as very hard), following the interpretative framework for difficulty levels as outlined by [Hambleton *et al.*, 1991; Baker and Kim, 2017; Giacomoni *et al.*, 2015].

It is noteworthy that, while the Bebras organizers consistently predicted tasks to be either easy or medium, their precision in precisely pinpointing the difficulty levels of individual items was notably inaccurate. Consequently, based on the IRT analysis and considering the 58% difficulty prediction precision, we can deduce that Instrument One falls within the easy to medium difficulty level range.

Table 10. Difficulty of Item by IRT and Bebras Organizers - Instrument One

Item	Tasks	IRT	Bebras
1	Drawing stars	Easy	Easy
2	Bowl Factory	Hard	Medium
3	Email	Very Easy	Easy
4	Beaver alchemist	Medium	Medium
5	Tutorial	Easy	Easy
6	Popularity	Medium	Medium
7	Word chain	Very Easy	Medium
8	Fireworks	Medium	Hard
9	You won't find it	Easy	Easy
10	Turn the cards	Very Hard	Easy
11	Decorating chocol.	Medium	Easy
12	Busy beaver	Hard	Hard

Table 11 outlines the difficulty classification for Instrument Two, incorporating assessments from both IRT and Bebras organizers. The breakdown reveals that six items are deemed easy tasks (with three of them falling under the very easy category), seven are categorized as medium, and two are considered hard (with none classified as very hard). It is

noteworthy that, despite the Bebras organizers' prediction of five items as difficult, their categorization proved to be inaccurate. Taking into account both IRT and organizers' classifications, the precision in difficulty level determination for Instrument Two stands at 53%.

This suggests that Instrument Two, based on the combined assessment, falls within the medium difficulty level range, although the precision in difficulty prediction is somewhat modest.

Table 11. Difficulty of Item by IRT and Bebras Organizers - Instrument Two

Item	Tasks	IRT	Bebras
1	Ceremony	Hard	Easy
2	Log-art	Very easy	Easy
3	Beavers on the run	Very easy	Easy
4	Traffic in the city	Easy	Easy
5	Storm proof network	Medium	Medium
6	Space maze	Very easy	Easy
7	Footprints	Easy	Medium
8	Puddle jumping	Easy	Medium
9	Social network	Medium	Medium
10	Height game	Medium	Hard
11	Meeting point	Medium	Hard
12	Best translation	Medium	Hard
13	Broken machines	Medium	Medium
14	True or false	Medium	Hard
15	Right rectangles	Hard	Hard

Predicting the difficulty level of questions is a multifaceted challenge influenced by various factors, including (i) the perspectives of stakeholders proposing the instrument, (ii) the implicit knowledge of test designers, (iii) the characteristics of the subjects undertaking the instrument, and (iv) intrinsic features of the questions [Van der Vegt, 2013; Dhillon, 2003]. Despite concerted efforts, achieving complete reliability in difficulty prediction remains elusive. A more effective approach involves examining test results rather than attempting to pre-determine difficulty levels [Baker and Kim, 2017; De Ayala, 2013; Hutz *et al.*, 2015].

Classical Test Theory (CTT) posits that the facility or difficulty level is gauged by the total correct or incorrect responses on an instrument. For instance, the facility rate is computed by dividing the number of correct responses by the total number of subjects, yielding a percentage ranging from 0 to 100%. A higher percentage indicates an easier question (above 75% is classified as easy), suggesting that a substantial number of examinees answered correctly. Conversely, a lower percentage (below 25%) categorizes a question as difficult, implying that only a few subjects endorsed the correct answer [Erthal, 1987]. The raw test score, representing the sum of scores on the items, provides an overall assessment of each student's performance on a standardized scale for the specific test [De Ayala, 2013].

In contrast, under IRT, the difficulty level is measured on the ability scale ( $\Theta$ ) when the probability of a correct response is 50% [Baker, 2001]. The Item Characteristic Curve (ICC) graphically illustrates the relationship between the probability of selecting the correct answer and the ability scale ( $\Theta$ ). For easy items, this point appears at a low ability level, while for difficult items, it occurs at a high ability level. As we adopt IRT, our focus on discussing difficulty level is based on the difficulty parameter rather than raw scores.

Our findings indicate that Bebras organizers achieved 53% to 58% accuracy in predicting difficulty levels. This un-

derscores the complexity of anticipating difficulty classifications. While in the context of the Bebras Challenge, where the primary objective is to promote computational thinking (CT), this may not pose a significant issue, our data suggest that if Bebras tasks are to be utilized as a measurement instrument, a more nuanced investigation is warranted.

### 5.3.1 Exploring Difficulty Levels

In delving into the examination of task difficulty, our focus extended to tasks with extreme difficulty values, encompassing those categorized as very easy, very hard, as well as those with critical difficulty parameter values. Within Instrument One, the tasks identified as the easiest include “Email,” “Word Chain,” “You Won’t Find It,” and “Tutorial.” Notably, “Email” and “Tutorial” tasks lack accompanying illustrations, distinguishing them as the only questions without visual aids in both instruments. Despite the absence of visual cues, these tasks are deemed easy due to their emphasis on evaluative abilities. Conversely, the “Word Chain” task relies heavily on visual perception, necessitating attention to pattern recognition in the illustration for a successful solution. Lastly, “You Won’t Find It” task demonstrates the application of a pre-defined algorithm in reverse order through a flowchart, with visual representation aiding algorithm application, requiring only a basic knowledge of alphabetical order.

Examining the easiest tasks in Instrument Two, we find “Log-art,” “Beavers on the Run,” and “Space Maze” tasks, all classified as very easy. Both “Log-art” and “Word Chain” tasks emphasize visual perception, prompting solvers to identify patterns and select alternatives that conform to the established pattern. In “Beavers on the Run,” an algorithm is introduced and applied in a more intricate scenario, while “Space Maze” involves associating alternative instructions with the directions a robot must follow to reach a goal. Despite requiring familiarity with cardinal directions, the task is still considered easy.

Conversely, the task classified as the hardest in Instrument One is “Turn the Card,” which demands logical reasoning for a solution. The initial impression might mislead solvers into thinking that only moving the vowel and the even card is required, making the statement a potential distractor. However, knowledge of the truth table increases the likelihood of selecting the correct answer—moving the vowel and old cards. Despite its initial complexity, this task is deemed inappropriate due to its critical difficulty parameter value.

In contrast, Instrument Two contains two hard tasks: “Ceremony” and “Right Rectangles.” “Ceremony” is labeled as very hard due to its free-response nature, with one correct answer. During test corrections, confusion arose as students mixed up numbers in sequence. A reevaluation of the question format, allowing solvers to provide the correct answer without potential distractions, is recommended. “Right Rectangles” is categorized as hard due to its inclusion of single and complex instructions, along with spatial and mathematical concepts such as “turn 90° clockwise”  $n$ -times.

In summary, the IRT approach to difficulty parameters as a location index sets it apart from the Classical Test Theory. Consequently, an instrument requires easy, medium,

and hard questions to cover the entire ability scale, enabling assessment across different proficiency levels. In our study, both instruments predominantly feature easy and medium tasks, with a limited number of hard questions. This indicates that both instruments may effectively assess individuals with low and medium computational thinking skills but might fall short in evaluating those with high abilities. Thus, our findings support the assertion that (i) predicting the difficulty classification of items in advance may lack accuracy, and (ii) IRT proves to be a suitable choice for determining difficulty levels.

## 5.4 (RQ2) - Utility of Bebras Tasks in CT Assessment: To what extent do Bebras tasks serve as effective tools in assessing CT skills, as measured by IRT metrics?

Recalling that discrimination gauges the proficiency-distinguishing ability of a question, and difficulty is associated with how an examinee responds to an item, we assess the efficacy of tasks in gauging CT skills based on these criteria. Both Instruments feature items spanning various discrimination and difficulty levels.

In Instrument One, tasks exhibiting high discrimination and medium difficulty include “Beaver the Alchemist” and “Decorating Chocolate.” The former is deciphered by tallying objects and deciphering transformation rules illustrated through a graph. The latter entails classic algorithmic thinking, presenting code in multiple-choice alternatives where the solver must select the incorrect option, necessitating knowledge of geometric angles for resolution.

Tasks with moderate discrimination values are exemplified in “Bowl Factory” and “Popularity.” The former employs the bubble sort algorithm to order bowls of different sizes, posing a question about the iterations needed for sequencing. This task demonstrates good discrimination at high ability levels. The latter involves counting direct and indirect nodes based on a graph relation, effectively combining descriptive text and illustrative relations.

In Instrument Two, tasks with high discrimination values are conducive to examinees with lower abilities (easy tasks). The remaining tasks showcase moderate discrimination values, categorized as medium difficulty. For instance, the “Storm-Proof Network” task presents a network topology where towers represent nodes, challenging the solver to identify the node disrupting the connection.

Tasks with low discrimination values coincide with those posing difficulty classification challenges. “Email” and “Turn the Card” tasks, labeled as problematic, exhibit critical values on difficulty parameters. The “Busy Beaver” task is characterized by unclear rules, a lack of choices, and a challenging open-ended nature. Additionally, the “Fireworks” task, a visual perception question, introduces a distractor, potentially misleading solvers into underestimating the number of possible arrangements systematically.

In summary, both Instruments feature items with low, moderate, and high discrimination values, as well as easy and medium tasks, with few hard questions. Consequently, these Instruments can evaluate individuals with varying ability lev-

els in CT skills. However, an instrument comprising items with moderate and high discrimination across easy, medium, and hard questions would offer a more comprehensive assessment of CT skills.

## 6 Observed practices

This section delves into the insights gained from the problem-solving process and its correlation with the psychometric properties identified through Item Response Theory (IRT) analysis. Throughout this exploration, discernible patterns emerged as we linked specific characteristics and abilities to the observed discrimination and difficulty level outcomes. Consequently, we respond to our third research question (RQ3 - Best Practices for Item Development in CT Assessment: What are the recommended practices for the development of items geared towards the assessment of CT, particularly within the framework of Bebras Challenge tasks?). The analysis of this question encompasses considerations on the design of CT items and instruments. Finally, we address potential threats to the validity of our study.

The results presented in this section were obtained using a qualitative methodology inspired by focus group techniques. A focus group is a research technique that enables data collection through group interactions when discussing a specific topic suggested by the lead researcher [Morgan, 1996]. At this stage of our research, we invited the Computer Science teachers who participated in the question translation process. During the focus group meetings, characteristics previously identified by the lead researcher were discussed, and patterns were observed and recorded. These records are described in this section.

### 6.1 Computational Thinking Items

Items serve as the fundamental components of any assessment instrument. Our analysis has revealed five discernible patterns related to the examination of tasks, taking into account difficulty and discrimination levels.

*1. Items with Figures Demonstrate Satisfactory Discrimination:* Most items incorporate illustrations, which serve various purposes such as explaining the problem, illustrating statements, or aiding in the understanding of steps. However, a general illustration without a clear purpose may not enhance discrimination and could potentially be distracting. Figures that contribute to problem comprehension or mental representation tend to increase discrimination. Therefore, illustrations should depict the problem or solution to effectively aid solvers.

*2. Items with Graphs and Text Rules Present Excellent Discrimination:* Graphs, as a type of data structure, are effective in showcasing relationships among elements, making it easier to visualize connections, algorithms, and rules. When graphs are linked to text rules, the redundancy in information arises, where the figure essentially represents the text rules. This redundancy allows motivated and careful solvers, or those who bypass the text description, to solve the problem equally well. Associating text rules with graphs is deemed essential for enhancing discrimination.

*3. Items with Logical Statements Require Careful Consideration for Good Discrimination:* Tasks featuring logical statements come in two types, one with medium difficulty and moderate discrimination, and the other with low discrimination and a very hard difficulty level. The key distinction lies in the fact that the former, exemplified by the “True and False” task, solely employs logical statements, while the latter, represented by the “Turn the Card” task, utilizes the truth table rule. The use of a specific truth table rule is considered suboptimal for designing items that do not demand prior knowledge. Designers should carefully consider the type of logical statements they employ to ensure effective discrimination.

*4. Items that Explore Algorithm Execution Skills with Commands Present Good Discrimination Across All Ability Levels:* Tasks with explicit commands requiring the execution of an algorithm cover the entire ability scale (Theta) and exhibit moderate to high discrimination values. Notable instances include tasks such as “Right Rectangles,” “Broken Machine,” “Decorating Chocolate,” “Footprints,” and “Space Maze”. Items involving algorithms with commands offer an effective option for designing questions with robust discrimination that spans the ability scale. The difficulty level is believed to be associated with the evaluation skill in this context.

*5. Recognizing Sub-Problems (Decomposition) Can Be Associated with Two Kinds of Solutions: Recursion or Establishing Intermediate Results:* Recognition of sub-problems (decomposition) can lead to either recursive solutions, where the previous answer systematically contributes to the subsequent response, or the establishment of intermediate results. For example, in the “Footprint” task, the final solution depends on answering smaller instances of the same problem statement. In “Height Game,” the solver must recognize two sub-problems: initially ordering characters by height and then applying restrictions to achieve the final solution. Designers should be mindful of the type of decomposition they request for problem resolution.

Finally, we address final considerations about critical issues that we observed. For example, the “Turn the Card” task introduces a distractor based on an initial impression that may mislead solvers, while the “Email” task lacks visual aids, potentially oversimplifying the cognitive process involved. Future refinements could involve redesigning these tasks to incorporate elements that demand a broader range of CT skills, such as providing additional contextual information or requiring multi-step reasoning (decomposition skill).

Related to tasks with unclear rules, we highlight the “Busy Beaver” task, characterized by ambiguous rules and a lack of well-defined choices, and the open-ended “Fireworks” task, which introduces potential distractors, highlights the challenges in task design. These issues might confuse solvers and reduce the reliability of their responses. To address this, future iterations of these tasks could adopt clearer and more structured formats. For instance, explicitly stating the rules and offering multiple-choice options can enhance the clarity and usability of the tasks, reducing the cognitive load on solvers.



## 6.2 Designing Computational Thinking Assessment Instruments

In the pursuit of creating reliable instruments to measure CT skills, we draw upon five key lessons learned from this research, derived from a comprehensive investigation into the assessment of CT skills and insights from psychometric literature [Baker and Kim, 2017; De Ayala, 2013; Hutz *et al.*, 2015; Pasquali, 2017].

1. *It is crucial to follow a consolidated methodology:* A robust methodology provides a secure framework guiding the production of a trustworthy instrument. It should detail all necessary steps, aligning with the research objective and the desired instrument type. In this context, IRT stands out as a robust and practical approach for designing CT assessment instruments. IRT parameters offer valuable insights into difficulty level, discrimination, and guessing by chance.

2. *The concept and associated CT skills explored in the assessment instrument should be as precise as possible:* Clear definitions of CT and the associated skills under exploration are fundamental. These specifications guide the design of items tailored to measure these skills. The abilities should be detailed, focusing on the actions required to solve the questions. Lack of clarity in these concepts during item design can lead to inaccuracies in the entire instrument.

3. *A group of CT specialists should evaluate and judge the content and the associated skills in the instrument:* Content and semantic analysis by CT specialists or trained raters is essential for ensuring the intelligibility of items. Specialists assess whether the problem statements can be easily understood, devoid of elements that may cause embarrassment or distraction. Furthermore, they verify whether the items effectively explore problems that require CT skills, identifying the associated skills employed during the resolution process.

4. *Predicting the difficulty level of items may not be accurate if we consider the organizers' tacit knowledge exclusively:* The perception of difficulty can vary depending on the classification strategy, and relying solely on organizers' knowledge may lead to underestimation or overestimation. Our study revealed approximately 58% accuracy when comparing difficulty predictions by Bebras' organizers (pre-application) with IRT results (post-application). Accurate assessment of difficulty is often better achieved after students have interacted with the items.

In addition, intentional or accidental difficulty in items should be considered. Designers must carefully plan for complexity, while unintended difficulty may only become apparent after item application. Our study observed that certain tasks involved elementary mathematics content, such as basic arithmetic and geometry angles. However, these elements did not pose the main difficulty in problem-solving. The study suggests that the mathematical content, although present, is elementary for the study participants and does not present a significant challenge. Designers should reconsider this point, especially when assessing children, as mathematical difficulty may be underestimated.

Lastly, measuring skills without relying on a total score of correct answers can benefit the CT field. Traditional approaches using overall scores have limitations, such as equating the ability level of individuals with the same total score,

even if they endorse items of different difficulty levels. To overcome these limitations, the IRT methodology proves valuable by allowing for the calibration of instruments with easy, medium, and hard items, offering a nuanced understanding of cognitive abilities involved in CT.

## 6.3 Limitations

Like any experimental research, our study is subject to certain limitations that may affect its internal and external validity. Given that our research involves cognitive assessment, human factors pose a potential threat to internal validity. Our participant selection process involved inviting all students enrolled in course to participate voluntarily, leading to a non-random sample. However, we implemented statistical procedures to ensure that the groups were comparable and drawn from the same population. While we engaged students from two universities, caution is warranted in generalizing our results to all undergraduate students. Nevertheless, we ensured a sufficient minimum sample size to maintain statistical rigor during IRT procedures. It is important to note that IRT parameters are not contingent on the sample size. Another limitation of our study is that it was conducted in 2017 and 2018, before the COVID-19 pandemic. As such, our research reflects a snapshot of that specific period.

Although translation processes aim to preserve the construct measured, in our case, computational thinking, cultural variations may interfere and introduce bias into the translated test. For instance, curiosity about the animal depicted in some questions — a beaver (an animal not found in Brazilian fauna) — might have served as a distraction. A solution for future studies is to adopt more rigorous translation techniques commonly used in Psychometrics for psychometric tests. These techniques involve engaging professionals in linguistics or language studies who are fluent in both the source and target languages, as well as native speakers of both languages. Additionally, the process includes cross-cultural adaptation, back-translation, and multiple pilot tests to ensure the instrument is reliable for measuring the intended skill.

At last, we highlighted that although we adopted a scoring system different from that proposed by the Bebras Challenge contest in this study, this did not affect the results regarding correct and incorrect responses. It is because our research focused on investigating the psychometric properties of the items rather than the total score of correct answers per student.

## 7 Conclusion

This study aimed to scrutinize the design of questions within the Bebras Challenge framework as instruments for measuring Computational Thinking (CT) skills. Employing Item Response Theory (IRT) metrics, we assessed the psychometric properties of Bebras tasks, focusing on difficulty and discrimination parameters.

Addressing our first research question (RQ1) on the accuracy of predicting difficulty levels, we found that the prediction was approximately 53% to 58% accurate when con-



fronted with empirical data. This suggests caution in advancing difficulty level classifications without relying on empirical evidence, highlighting the need for precise and data-driven approaches.

Regarding RQ2, examining the effectiveness of Bebras tasks in assessing CT skills through IRT metrics, we observed that Bebras tasks predominantly exhibit easy and medium difficulty levels. The discrimination values varied, with items displaying low, moderate, and high discrimination. This finding underscores the importance of rethinking question design, emphasizing the necessity of creating tasks across all difficulty levels with robust discrimination values to ensure comprehensive proficiency assessment.

In addressing our third research question (RQ3) and drawing lessons from the study, we propose recommended practices for developing items tailored for the assessment of CT within the Bebras Challenge framework. These practices emerged from patterns associating characteristics and abilities with discrimination and difficulty levels. We advocate for the adoption of a consolidated methodology, precision in defining CT concepts and skills, and the involvement of CT experts in item evaluation.

Looking forward, potential avenues for future research involve evaluating additional item features, such as vocabulary, text size, and the number of elements or steps needed to reach a solution. Qualitative studies on students' perceptions of difficulty could provide valuable insights. Moreover, investigating the impact of item modifications on discrimination and difficulty levels could offer valuable refinements to CT assessment instruments. In conclusion, we hope this study sparks further exploration into the quantification of CT skills.

## Declarations

## Acknowledgements

The authors would like to thank the research participants.

## Funding

This research was funded by the authors.

## Authors' Contributions

All authors were involved in every stage of the research. ALSO performed the data collection. All authors collaborated on the discussion of data analysis and writing the final article.

## Competing interests

The authors declare that they have no competing interests.

## Availability of data and materials

Data are accessible upon request from the corresponding author.

## References

- Baker, F. B. (2001). *The basics of item response theory*. ERIC. Book.
- Baker, F. B. and Kim, S.-H. (2017). *The basics of item response theory using R*. Springer. DOI: 10.1007/978-3-319-54205-8.
- Barr, V. and Stephenson, C. (2011). Bringing computational thinking to k-12: what is involved and what is the role of the computer science education community? *Acm Inroads*, 2(1):48–54. DOI: 10.1145/1929887.1929905.
- Basu, S., Rutstein, D., Xu, Y., and Shear, L. (2020). A principled approach to designing a computational thinking practices assessment for early grades. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, pages 912–918. DOI: 10.1145/3328778.3366849.
- Bellettini, C., Lonati, V., Malchiodi, D., Monga, M., Morpurgo, A., and Torelli, M. (2015). How challenging are bebras tasks?: an irt analysis based on the performance of italian students. In *Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education*, pages 27–32. ACM. DOI: 10.1145/2729094.2742603.
- Boom, K.-D., Bower, M., Arguel, A., Siemon, J., and Scholkmann, A. (2018). Relationship between computational thinking and a measure of intelligence as a general problem-solving ability. In *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, pages 206–211. ACM. DOI: 10.1145/3197091.3197104.
- Brennan, K. and Resnick, M. (2012). New frameworks for studying and assessing the development of computational thinking. In *Proceedings of the 2012 annual meeting of the American Educational Research Association, Vancouver, Canada*. Available at: <https://scratched.gse.harvard.edu/ct/files/AERA2012.pdf>.
- Csizmadia, A., Curzon, P., Dorling, M., Humphreys, S., Ng, T., Selby, C., and Woollard, J. (2015). Computational thinking - a guide for teachers. Project report, University of Southampton Institutional Repository. Available at: <https://eprints.soton.ac.uk/424545/>.
- Dagienė, V. and Futschek, G. (2008). Bebras international contest on informatics and computer literacy: Criteria for good tasks. In *International Conference on Informatics in Secondary Schools-Evolution and Perspectives*, pages 19–30. Springer. DOI: 10.1007/978-3-540-69924-8\_2.
- Dagiene, V., Mannila, L., Poranen, T., Rolandsson, L., and Söderhjelm, P. (2014). Students' performance on programming-related tasks in an informatics contest in finland, sweden and lithuania. In *Proceedings of the 2014 conference on Innovation & Technology in Computer Science education*, pages 153–158. ACM. DOI: 10.1145/2591708.2591760.
- Dagienė, V., Pelikis, E., and Stupurienė, G. (2015). Introducing computational thinking through a contest on informatics: Problem-solving and gender issues. *Informacijos Mokslai/Information Sciences*, 73. Available at: <https://www.ceeol.com/search/>

- article-detail?id=467909.
- Dagienė, V., Sentance, S., and Stupurienė, G. (2017). Developing a two-dimensional categorization system for educational tasks in informatics. *Informatica*, 28(1):23–44. DOI: 10.15388/Informatica.2017.119.
- Dagienė, V., Stupurienė, G., et al. (2016). Bebras-a sustainable community building model for the concept based learning of informatics and computational thinking. *Informatics in Education-An International Journal*, 15:25–44. Available at: <https://eric.ed.gov/?id=EJ1097494>.
- De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications. Book.
- Dhillon, D. (2003). Predictive models of question difficulty: A critical review of the literature. In *Technical Report. AQA Centre for Education Research and Policy, Manchester, UK*. Available at: [https://filestore.aqa.org.uk/content/research/CERP-RP-DD-01022003\\_0.pdf](https://filestore.aqa.org.uk/content/research/CERP-RP-DD-01022003_0.pdf).
- Dolgopolas, V., Jevsikova, T., Savulionienė, L., and Dagienė, V. (2015). On evaluation of computational thinking of software engineering novice students. In *Proceedings of the IFIP TC3 Working Conference "A New Culture of Learning: Computing and next Generations"*, pages 90–99. DOI: 10.13140/RG.2.1.2855.9206.
- Duncan, C. and Bell, T. (2015). A pilot computer science and programming course for primary school students. In *Proceedings of the Workshop in Primary and Secondary Computing Education*, pages 39–48. ACM. DOI: 10.1145/2818314.2818328.
- Erthal, T. C. (1987). *Manual de psicometria*. Zahar. Book.
- Giacomoni, C. H., de Lima Athayde, M., Zanon, C., and Stein, L. M. (2015). Teste do desempenho escolar: evidências de validade do subteste de escrita. *Psico-USF*, 20(1):133–140. DOI: 10.1590/1413-82712015200112.
- Giordano, D. and Maiorana, F. (2014). Use of cutting edge educational tools for an initial programming course. In *Global Engineering Education Conference (EDUCON), 2014 IEEE*, pages 556–563. IEEE. DOI: 10.1109/EDUCON.2014.6826147.
- Grover, S. and Pea, R. (2013). Computational thinking in k–12 a review of the state of the field. *Educational Researcher*, 42(1):38–43. DOI: 10.3102/0013189X124630.
- Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of item response theory*, volume 2. Sage. Book.
- Hubwieser, P. and Mühling, A. (2015). Investigating the psychometric structure of bebras contest: towards measuring computational thinking skills. In *Learning and Teaching in Computing and Engineering (LaTiCE), 2015 International Conference on*, pages 62–69. IEEE. DOI: 10.1109/LaTiCE.2015.19.
- Hutz, C. S., Bandeira, D. R., and Trentini, C. M. (2015). *Psicometria*. Artmed Editora. Book.
- Ilic, U., Haseski, H. İ., and Tugtekin, U. (2018). Publication trends over 10 years of computational thinking research. *Contemporary Educational Technology*, 9(2):131–153. DOI: 10.30935/cet.414798.
- ISTE (2011). Operational definition of computational thinking for k-12 education. Available at: <http://www.iste.org/docs/ct-documents/computational-thinking-operational-definition-flyer.pdf?sfvrsn=2>.
- Kalelioglu, F., Gülbahar, Y., and Kukul, V. (2016). A framework for computational thinking based on a systematic research review. *Baltic Journal of Modern Computing*, 4(3):583. Available at: <https://shorturl.at/cKlet>.
- Korkmaz, Ö., Çakir, R., and Özden, M. Y. (2017). A validity and reliability study of the computational thinking scales (cts). *Computers in Human Behavior*. DOI: 10.1016/j.chb.2017.01.005.
- Lockwood, J. and Mooney, A. (2017). Computational thinking in education: Where does it fit? a systematic literary review. *Maynooth University*. DOI: 10.48550/arXiv.1703.07659.
- Lockwood, J. and Mooney, A. (2018). Developing a computational thinking test using bebras problems. *European Conference on Technology Enhanced Learning - EC-TEL 2018*. Available at: [https://ceur-ws.org/Vol-2190/TACKLE\\_2018\\_paper\\_1.pdf](https://ceur-ws.org/Vol-2190/TACKLE_2018_paper_1.pdf).
- Lonati, V., Malchiodi, D., Monga, M., and Morpurgo, A. (2017). Bebras as a teaching resource: classifying the tasks corpus using computational thinking skills. In *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education*, pages 366–366. ACM. DOI: 10.1145/3059009.3072987.
- Mannila, L., Dagienė, V., Demo, B., Grgurina, N., Mirolo, C., Rolandsson, L., and Settle, A. (2014). Computational thinking in k-9 education. In *Proceedings of the working group reports of the 2014 on innovation & technology in computer science education conference*, pages 1–29. ACM. DOI: 10.1145/2713609.2713610.
- Matsuzawa, Y., Murata, K., and Tani, S. (2018). Multivocal challenge toward measuring computational thinking. In *Open Conference on Computers in Education*, pages 56–66. Springer. DOI: 10.1007/978-3-030-23513-0\_6.
- Mooney, A. and Lockwood, J. (2020). The analysis of a novel computational thinking test in a first year undergraduate computer science course. *AISHE-J: The All Ireland Journal of Teaching and Learning in Higher Education*, 12(1):1–27. DOI: 10.62707/aishej.v12i1.420.
- Moreno-León, J., Román-González, M., and Robles, G. (2018). On computational thinking as a universal skill: A review of the latest research on this ability. In *2018 IEEE Global Engineering Education Conference (EDUCON)*, pages 1684–1689. IEEE. DOI: 10.1109/EDUCON.2018.8363437.
- Morgan, D. L. (1996). *Focus groups as qualitative research*, volume 16. Sage publications. DOI: 10.4135/9781412984287.
- Palts, T., Pedaste, M., Vene, V., and Vinikienė, L. (2017). Tasks for assessing skills of computational thinking. In *ICERI2017 Proceedings*, 10th annual International Conference of Education, Research and Innovation, pages 2750–2759. IATED. DOI: 10.1145/3059009.3072999.
- Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. Basic Books, Inc. Book.
- Papert, S. and Harel, I. (1991). Situating constructionism. *Constructionism*, 36(2):1–11. Available at: <https://www.iste.org/docs/ct-documents/computational-thinking-operational-definition-flyer.pdf?sfvrsn=2>.

- //web.media.mit.edu/~calla/web\_comunidad/Reading-En/situating\_constructionism.pdf.
- Pasquali, L. (2017). *Psicometria: teoria dos testes na psicologia e na educação*. Editora Vozes Limitada. Book.
- Pohl, W. and Hein, H.-W. (2015). Aspects of quality in the presentation of informatics challenge tasks. In *The Proceedings of International Conference on Informatics in Schools: Situation, Evolution and Perspectives—ISSEP 2015*, page 21. Available at: <https://issep15.fri.uni-lj.si/files/issep2015-proceedings.pdf>.
- Román-González, M. (2015). Computational thinking test: Design guidelines and content validation. In *Proceedings of the 7th Annual International Conference on Education and New Learning Technologies (EDULEARN 2015)*, pages 2436–2444. Available at: <https://library.iated.org/view/ROMANGONZALEZ2015COM>.
- Román-González, M., Pérez-González, J.-C., and Jiménez-Fernández, C. (2016). Which cognitive abilities underlie computational thinking? criterion validity of the computational thinking test. *Computers in Human Behavior*. DOI: 10.1016/j.chb.2016.08.047.
- Seiter, L. and Foreman, B. (2013). Modeling the learning progressions of computational thinking of primary grade students. In *Proceedings of the ninth annual international ACM conference on International computing education research*, pages 59–66. ACM. DOI: 10.1145/2493394.2493403.
- Selby, C. C. (2014). *How can the teaching of programming be used to enhance computational thinking skills?* PhD thesis, University of Southampton. 325pp. Available at: <https://eprints.soton.ac.uk/366256/>.
- Selby, C. C. (2015). Relationships: Computational thinking, pedagogy of programming, and bloom’s taxonomy. In *Proceedings of the Workshop in Primary and Secondary Computing Education, WiPSCE ’15*, pages 80–87, New York, NY, USA. ACM. DOI: 10.1145/2818314.2818315.
- Sherman, M. and Martin, F. (2015). The assessment of mobile computational thinking. *Journal of Computing Sciences in Colleges*, 30(6):53–59. DOI: 10.5555/2753024.2753037.
- Tang, X., Yin, Y., Lin, Q., Hadad, R., and Zhai, X. (2020). Assessing computational thinking: A systematic review of empirical studies. *Computers & Education*, 148:103798. DOI: 10.1016/j.compedu.2019.103798.
- Tikva, C. and Tambouris, E. (2021). Mapping computational thinking through programming in k-12 education: A conceptual model based on a systematic literature review. *Computers & Education*, 162:104083. DOI: 10.1016/j.compedu.2020.104083.
- Tsarava, K., Moeller, K., Román-González, M., Golle, J., Leifheit, L., Butz, M. V., and Ninaus, M. (2021). A cognitive definition of computational thinking in primary education. *Computers & Education*, page 104425. DOI: 10.1016/j.compedu.2021.104425.
- Van der Vegt, W. (2013). Predicting the difficulty level of a bebras task. *Olympiads in Informatics*, 7:132–139. Available at: <https://ioinformatics.org/journal/INFOL127.pdf>.
- Van der Vegt, W. (2018). How hard will this task be? developments in analyzing and predicting question difficulty in the bebras challenge. *Olympiads in Informatics*, 12:119–132. Available at: [https://ioinformatics.org/journal/v12\\_2018\\_119\\_132.pdf](https://ioinformatics.org/journal/v12_2018_119_132.pdf).
- VANÍČEK, J., ŠIMANDL, V., and KLOFÁČ, P. (2021). A comparison of abstraction and algorithmic tasks used in bebras challenge. *Informatics in Education*, 20(4):717. DOI: 10.15388/infedu.2021.30.
- Vendramini, C. M. M. and Dias, A. S. (2005). Teoria de resposta ao item na análise de uma prova de estatística em universitários. *Psico-USF*, 10(2):201–210. DOI: 10.1590/S1413-82712005000200012.
- Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, 49(3):33–35. DOI: 10.1145/1118178.1118215.
- Wing, J. M. (2008). Computational thinking and thinking about computing. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 366(1881):3717–3725. DOI: 10.1098/rsta.2008.0118.
- Yagunova, E., Podznyakov, S., Ryzhova, N., Razumovskaia, E., and Korovkin, N. (2015). Tasks classification and age differences in task perception. case study of international on-line competition “beaver”. In *Proc. of the 8th ISSEP Conf*, pages 33–43. Available at: <https://core.ac.uk/download/pdf/77923211.pdf#page=39>.