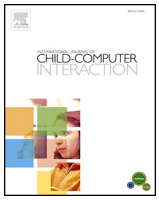
International Journal of Child-Computer Interaction 40 (2024) 100650

Contents lists available at ScienceDirect

International Journal of Child-Computer Interaction

journal homepage: www.elsevier.com/locate/ijcci



Research Paper

Validating a performance assessment of computational thinking for early childhood using item response theory

Chungsoo Na [a], Jody Clarke-Midura [a,*], Jessica Shumway [b], Wilhelmina van Dijk [c], Victor R. Lee [d]

[a] 2830 Old Main Hill, Utah State University, Logan, UT, 84322-2830, USA
[b] 2805 Old Main Hill, Utah State University, Logan, UT, 84322-2830, USA
[c] 2865 Old Main Hill, Utah State University, Logan, UT, 84322-2865, USA
[d] 485 Lasuen Mall, Stanford University, Stanford, CA, 94305, USA

ARTICLE INFO

Keywords:
Early childhood
Computational thinking
Item response theory
Assessment
Evidence-centered design.

ABSTRACT

Despite growing interest in early childhood computational thinking (CT), there is a lack of validated assessments for children who are emerging readers. This paper presents validity and reliability evidence of a performance-based assessment of CT using item response theory (IRT) from 272 children aged 4–8. Using a two-parameter logistic model IRT model (2PL IRT), we confirmed that model- and item-level fits are acceptable. Item analyses revealed a *high* discriminability ($M = 2.26$, $SD = 1.12$) and a *moderate* item difficulty ($M = -0.21$; $SD = 0.86$), on average, across 19 items. Reliability analysis demonstrated that the assessment was substantially reliable (marginal reliability: $r_{xx} = 0.87$). Differential item functioning (DIF) analyses indicated that the assessment estimated children's item parameters fairly, regardless of their gender and age. However, we confirmed gaps in latent ability (θ) of CT by gender and age: boys showed higher latent ability of CT than girls, and old children (above 72 months) showed higher latent ability than young children (below 72 months). Findings suggest the assessment is a fair measure that can serve as a reliable and valid tool to assess CT for children who are emerging readers.

1. Introduction

The increase in access to tangible coding toys and applications like Scratch Jr have created more opportunities to introduce coding and computational thinking (CT) in early childhood classrooms (Bers, 2018; Zeng et al., 2023). Coding is a common context for developing CT, and young children can learn to program tangible coding toys and engage in CT before they can read and write (Bers, 2018; Relkin et al., 2020; Wang et al., 2021). Such tangible coding environments designed for young children who are pre-literate or emerging readers rely on symbol systems to represent codes to program an agent's movement or actions. Thus, young children learn to sequence using codes such as forward, backwards, rotate right, and rotate left. Examples of coding toys and the codes that correspond to movement are presented in Fig. 1.

Despite the increase in access to coding toys and apps, there is not an agreed upon definition of early childhood CT nor is there agreement on how to assess it (Clarke-Midura et al., 2021; 2023;Su & Yang, 2023). This is due to the novelty of the field of early childhood CT (Bers, 2018; Relkin et al., 2020; Zeng et al., 2023). In a recent review that looked at how CT is integrated in early childhood, Su and Yang (2023) identified four challenges for early childhood CT: the need for deeper learning of CT, lack of valid and reliable assessments, selecting developmentally appropriate tools, and developmentally appropriate curriculum. Other researchers have also indicated the need for validated and reliable assessments for early childhood CT (Clarke-Midura et al., 2021; Relkin et al., 2020; Tang et al., 2020). In addition, Su and Yang (2023) found that existing assessments of early childhood CT measure a range of concepts from programming skills to CT skills.

Given the lack of an agreed upon definition of early childhood CT, as part of a larger project, we operationalized a cognitive model of early childhood CT and developed research-based curricula tasks around tangible coding toys for kindergarten classrooms (Clarke-Midura et al., 2023; Shumway et al., 2023). We then developed a performance-based assessment we call CaST, which stands for computational and spatial

* Corresponding author. Department of Instructional Technology and Learning Sciences, College of Education and Human Services, Utah State University, 2830 Old Main Hill, Logan, UT, 84322-2830, USA.
E-mail address: jody.clarke@usu.edu (J. Clarke-Midura).

https://doi.org/10.1016/j.ijcci.2024.100650
Received 25 May 2023; Received in revised form 24 February 2024; Accepted 12 March 2024
Available online 14 March 2024

thinking assessment (Clarke-Midura et al., 2021). Our intent was to design a developmentally appropriate assessment that did not rely on children's reading or writing abilities and that could be used across tangible coding toy contexts. As part of our commitment to assessment fairness (Davidson et al., 2021), and knowing that some studies reported differences in CT knowledge based on participant's gender and age (Macrides et al., 2022), we want to test that the probabilities of answering assessment tasks correctly are attributed to children's true abilities in CT and do not favor or benefit participants based on their gender or age. Our research questions are: 1) *Is the CaST assessment a reliable and valid measure of CT for a sample of children aged 4–8? 2) Does the CaST assessment function equally for children regardless of their gender and age?*

The present study contributes to the current discourse on the challenges facing early childhood CT by providing evidence of the validity, reliability, and fairness of CaST with a sample of 272 children. In the sections that follow, we first discuss how CT is defined in early childhood and our operational definition of CT for early childhood. Next, we describe the research on assessments of early childhood CT that has reported validity evidence. We then discuss our methods and materials, including participants, data, and data analysis. Finally, we present our results followed by our discussion and conclusion.

2. Literature review

2.1. Computational thinking in early childhood

While the ideas behind computational thinking (CT) have roots in Papert's seminal work around LOGO (Papert, 1980), the term *computational thinking* was popularized by Jeannette Wing in her 2006 article in which she described it as a skill that involves "solving problems, designing systems, and understanding human behavior, by drawing on the concepts fundamental to computer science" (Wing, 2006, p. 33). Since then, researchers have developed various frameworks for defining CT, often shaped by particular context or for a particular age group (Angeli & Valanides, 2020; Bers, 2018; Brennan & Resnick, 2012; Martins et al., 2023; Wang et al., 2023; Zeng et al., 2023). Yet, as mentioned previously, there is not an agreed upon definition or framework for early childhood CT. Bers (2018) writes about CT in terms of powerful ideas, in which she proposed seven developmentally appropriate ideas for early childhood CT: algorithms, modularity, control structures, representation, hardware/software, design process, and debugging (Bers, 2018). Other researchers in early childhood have looked at sequencing (Angeli & Valanides, 2020; Città et al., 2019), debugging (Heikkilä & Mannila, 2018), and decomposition (Rijke et al., 2018). Thus, we started with these CT concepts and then spent hours in kindergarten classrooms observing the kinds of skills children used when they engaged with tangible coding toys (Clarke-Midura et al., 2021, 2023, Shumway et al., 2023). Our classroom studies resulted in what we refer to as a cognitive model of early childhood CT. For the design of the assessment, we only focused on the skills we knew we could observe and measure. Our cognitive model includes CT concepts such as algorithmic thinking, decomposition (modularity), debugging, and abstraction (Clarke-Midura et al., 2021, 2023). We also identified pre-requisite spatial and mathematical thinking knowledge that children used when they played with the tangible coding toys. For example, children reason with an agent's orientation, location, and navigation in space. We refer to these as foundational ideas and math knowledge (Clarke-Midura et al., 2021, Shumway et al., 2023). Table 1 lists the CT components of our model with definitions and Table 2 lists the pre-requisite spatial and mathematical thinking knowledge.

While most models of CT do not contain spatial thinking concepts, many researchers have investigated the relationship between CT and spatial ability. For example, in a sample with 1251 students in 5-10th grade, Román-González et al. (2017) found a significantly positive correlation between spatial ability and CT ($r = 0.44$, $p < 0.01$). This result was replicated by Tsarava et al. (2022) who reported a modest, positive association between visuospatial abilities and CT ($r = 0.35$, $p < 0.001$) in 192 3–4th primary school students. Similarly, Città et al. (2019) found that mental rotation ability was a significant predictor of CT skills for students in both grades 1–2 ($\beta = 2.13$; $p = 0.02$) and grades 3–6 ($\beta = 2.37$; $p < 0.001$). While these studies do not specifically investigate the co-occurrence of CT and spatial thinking, they support our position that children use spatial thinking when playing with coding toys and our decision to include it in our model as a *foundational idea*.

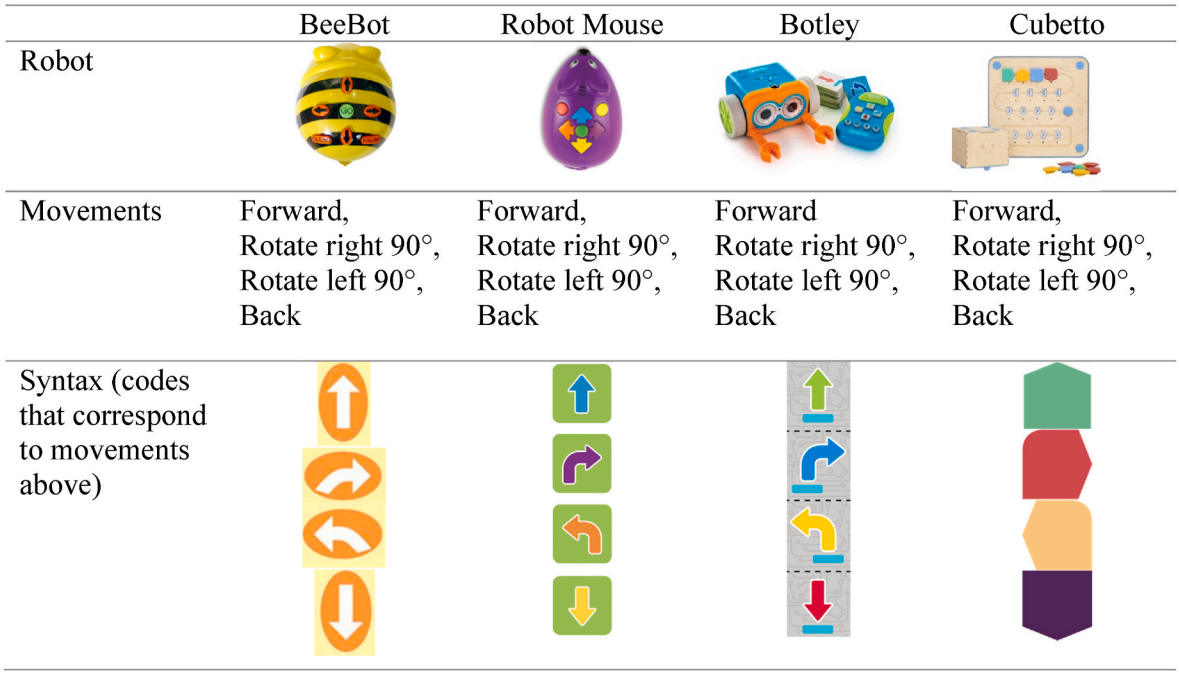


Fig. 1. Examples of Coding Toy for Early Childhood Computational Thinking.

Table 1
Operational Definitions of Components of our Early Childhood CT Cognitive Model.

CT Component	Operationalized definition
Algorithm thinking	Involves developing and using ordered sequences of instructions. Important subcomponents of algorithmic thinking are: • Sequencing codes- Ordering and arranging codes based on knowledge of syntax and semantics • Planning programs- Ordering and arranging codes based on knowledge of syntax and semantics • Reading/enacting programs- Interpreting (reading) and executing (enacting) sequence of codes
Debugging	Involves recognizing bugs/errors exist, locating the specific error or bug, proposing a fix, and correcting the bug. Important subcomponents of debugging are: • Recognizing Bug- Noticing that instructions do not work as expected or desired, or anticipating a problem before executing the program (i.e. knowing that there is a bug) • Locating Bug- Finding the part in the program that caused the problem (i.e. knowing where the bug is) • Proposing Solution- Making a plan or suggestion for how the program could change (i.e., knowing how to fix it) • Fixing Bug- Implementing a successful repair strategy (i.e. resolving the bug)
Decomposition	Involves recognizing parts in part-whole relationships, building a whole from parts, and breaking a whole into parts. Important subcomponents of decomposition are: • Breaking whole into parts- Recognize how whole programs can be broken down into units or segments of code to simplify the task/problem • Building whole from parts- Writing program by combining chunks or sequencing codes one-by-one • Relating parts to whole- Coordinating units or segments of code with one another as well as with whole program

Table 2
Foundational Ideas and Math Knowledge that are Pre-requisites for Solving CT Tasks.

Pre-requisite Spatial and Math Knowledge	Operationalized definition
Space-symbol coordination	Knowing how codes or parts of programs correspond to movements or paths traveled by the agent.
Spatial orientation	Knowing that the codes always produce the same movements but depend on the agent's orientation.
Spatial Code Meanings	Knowing what each of the codes instructs the agent to do.
One Code to One Movement Correspondence	Knowing that one code produces a single discrete linear or rotational movement.
Spatial reasoning	Knowing how the agent moves in 3 dimensions and thinking about them in different positions and orientations.
Counting on	Knowing that one code produces a single discrete linear or rotational movement.
Sequencing	Knowing that you do not include the starting location when counting forward movements.
Linear Units	Knowing how to use a standard unit of measure to make the agent travel along a linear path.
Rotation on a point	Knowing that an agent's rotation occurs by rotating on a fixed point at a set angle, not translating to an adjacent point.

2.2. Existing valid and reliable assessments of early childhood CT

As mentioned previously, there is a need for validated and reliable assessments of early childhood CT (Zeng et al., 2023). As shown in Table 3, we identified seven assessments of early childhood CT (from nine empirical studies) that reported reliability and validity evidence.

TechCheck is a multiple-choice assessment with 15 items (Relkin et al., 2020), designed to assess Bers' powerful ideas of CT in early childhood (Bers, 2018). In order to gather validity and reliability evidence, they conducted a study with 768 children ages 5-to-9. They reported an acceptable reliability ($\alpha = .68$), and appropriate item discrimination ($M = 1.03$) and low level of item difficulty ($M = -1.25$), on average, for their targeted population. However, the test information function of TechCheck peaked at low latent ability (θ); indicating this assessment is better at differentiating between children with relatively low CT ability (Relkin et al., 2020).

TechCheck-K is a modified version of TechCheck designed specifically for kindergarten-aged children (Relkin & Bers, 2021). This assessment is akin to TechCheck in terms of the multiple-choice items and constructs measured, however, to make it developmentally appropriate for kindergarten children and to adjust item difficulties, they reduced the number of possible answer choices from 4 to 3. They conducted a study with 87 children in which the item correction patterns of TechCheck-K were correlated with the patterns from the TechCheck ($r = 0.76$), suggesting that TechCheck-K can assess CT concepts in a comparable way to TechCheck. However, the authors presented relatively weak level of reliability and validity evidence of the TechCheck-K.

Similarly, Zapata-Cáceres et al. (2020) modified their Computational Thinking test (CTt, Román-González et al., 2017), for students aged 5 to 12, that they call the Beginners Computational Thinking Test (BCTt, Zapata-Cáceres et al., 2020). BCTt is 25 multiple-choice item assessment designed to measure sequences, loops, and conditionals. In a pilot study with 289 primary students, the BCTt showed good reliability (internal reliability: $\alpha = 0.82$; test-retest reliability: $r = 0.93$) and is more suitable for students in lower grades (i.e., 1–2 graders) than upper grades (3–6 graders) based on students' item correction patterns. This assessment was also validated by El-Hamamsy et al. (2022) with 374 3–4 graders using classical test theory (CTT) and item response theory (IRT). After excluding two items that were misfitted from the original 25 items, the results of IRT analyses with 23 items showed *moderate* item discrimination ($M = 1.58$) and relatively *easy* item difficulty ($M = -1.57$), on average. The assessment was easier for students in grades 3–4 and the authors claim that BCTt is more effective to use as a diagnostic tool to "discriminate between students with low abilities in grades 3 and 4" (El-Hamamsy et al., 2022, p. 17).

TACTIC-KIBO is an assessment for children aged 4–7 that is specific to the KIBO robot (Relkin et al., 2019). According to the children's coding ability, the difficulty and complexity of the assessment tasks gradually escalated from Level 1 to 4. This assessment was validated by Sung (Sung, 2022) with 450 Korean children aged 5–6 years, using IRT, CTT, and criterion validity with the Bebras Challenge (www.bebras.org) and early numeracy tasks (Howard & Melhuish, 2017). According to IRT results, TACTIC-KIBO had moderate mean item discriminations ($M = 1.77$), and item difficulties gradually increased with levels. However, although the CFA model of TACTIC-KIBO yielded acceptable fit indices, some items related to a specific platform (i.e., KIBO) or its specific functions had very weak factor loadings. Additionally, this knowledge did not correlate with the sub-factors of the Bebras tasks.

The Coding Stages Assessment (CSA) is an interview-based assessment of Bers' coding stages framework (de Ruiter & Bers, 2022). It consists of 27 items that ask children to answer questions verbally or to perform given tasks in ScratchJr. It is administered one-on-one by an administrator, who observes how the child performs a given task to determine whether he or she answers it correctly or not. Based on item responses from 118 children (5–8 years old), visual inspection of item characteristic curves from IRT analyses revealed good psychometric properties, and gender and age-related item bias was not detected. A moderate correlation ($r = 0.55$, $p < 0.05$) was found between CSA-ScratchJr and TechCheck scores on CT, indicating that the two assessments assessed the same CT construct. Since this assessment

Table 3
Synthesis of Assessment and Sample Features, Targeted CT components, and Validity and Reliability Evidence of Seven Early Childhood CT Assessments.

Validated measures	Assessment Format	Sample		Targeted CT components	Psychometric evidence	
		Age	*n*		Validity	Reliability
TechCheck (Relkin et al., 2020)	15 unplugged multiple-choice items	5–9	768	Algorithms, modularity, control structure, representation, debugging, and hardware/software	Face validity, Psychometric analysis (IRT) and Criterion validity via correlation with TACTIC-KIBO ($r = 0.54$)	Inter-rater reliability (Fleiss's $\kappa = 0.63$) and Internal reliability (Cronbach's $\alpha = 0.68$)
TechCheck-K (Relkin & Bers, 2021)	15 unplugged multiple-choice items	5–6	89	Algorithms, modularity, control structure, representation, debugging, and hardware/software	Correlations of the correct response patterns of the TechCheck-K and the TechCheck ($r = 0.76$)	Not provided in the manuscript
Beginners Computational Thinking test (BCTt) (Zapata-Cáceres et al., 2020)	25 unplugged-multiple choice items	5–12	299	Sequence, loop, conditional statements, and while statements	Face validity and Psychometric analysis (CTT)	Internal reliability (Cronbach's $\alpha = 0.82$) and test-retest reliability ($r = 0.93$ from 28 children)
Beginners Computational Thinking test (BCTt) (El-Hamamsy et al., 2022)	23 unplugged-multiple choice items	Grade 3-4	374	Sequence, loop, conditional statements, and while statements	Psychometric analysis (CTT and IRT)	Internal reliability (Cronbach's $\alpha = 0.82$ and Marginal reliability $r_{xx} = 0.75$)
TACTIC-KIBO (Relkin et al., 2019)	28 tasks (7 tasks × 4 levels)	5–7	15	Algorithms, modularity, control structure, representation, debugging, and hardware/software	Face validity and criterion validity via correlations with Interactive play sessions (IPS, $r = 0.90$)	Inter-rater reliability by experts
TACTIC-KIBO (Sung, 2022)	28 tasks (7 tasks × 4 levels)	5–6	108 (Level 4) – 332 (Level 1)	Control structure, hardware, software, representation, algorithms, modularity, debugging and design process	Psychometric analysis (IRT and CTT) and criterion validity via correlations with Bebras tasks ($r = 0.18$) and early numeracy ability ($r = 0.35$)	Internal reliability (Cronbach's $\alpha = 0.88$)
ScratchJr Coding Stage Assessment (CSA ScratchJr) (de Ruiter & Bers, 2022)	27 computer-based open-ended tasks	5–8	118	Emergent, coding and decoding, fluency, new Knowledge, and purposefulness	Construct validity, Criterion validity via correlations with TechCheck (from 23 children, $r = 0.55$), and Psychometric analysis (CTT, IRT and DIF)	Internal reliability (Guttman's $\lambda_6 = 0.94$) and Inter-rater reliability (Cohen's $\kappa = 0.78$)
KIBO project rubric (Govind & Bers, 2021)	Rubric	Grade 2	173 projects	1st iteration: General and KIBO-specific programming skills 2nd iteration: Programming concepts and project design	Face and construct validity	Inter-rater reliability (Cohen's weighted $\kappa = 0.84$)
ScratchJr Project rubric (Unahalekhaka & Bers, 2022)	Rubric	6–7	87 (228 projects)	Coding concepts and Project design	Face and construct validity and criterion validity via correlation with CSA-ScratchJr (de Ruiter & Bers, 2022; partial $r = 0.35$)	Inter-rater reliability (Krippendorff's $\alpha = 0.95$)

Note. CTT refers to classical test theory; IRT refers to item response theory; DIF refers to differential item functioning.

involves open-ended tasks, it is relatively long to administer ($M = 50$ min) and has wide variations in task-solving processes between children, both of which make it difficult to use in real-world settings.

KIBO Project Rubric (Govind & Bers, 2021) is a rubric-based assessment to measure KIBO projects in terms of programming concepts and project design elements. A score is awarded that provides an estimated level of mastery. Similarly, the ScractchJr Project Rubric (Unahalekhaka & Bers, 2022) evaluates ScratchJr projects based on coding concepts and project design. The rubric was validated in a study with 87 children aged 6–7 years. The inter-rater reliability of the rubric was high (Krippendorf's $\alpha = 0.95$). Criterion validity was verified through positive correlation ($r = 0.35$) between ScratchJr project rubric scores and the Coding Stage Assessment (CSA-ScratchJr) scores (de Ruiter & Bers, 2022), controlling for the effect of children's gender and grade levels.

While these assessments provide evidence of reliability and validity through psychometric analyses, some limitations of these assessments were identified: (a) some are tied to specific coding language or platform, (b) some rely on multiple-choice format, and (c) some assessments only work for children with relatively low CT abilities. In this regard, there is a need for a validated assessment that can be used across a variety of tangible coding toys and platforms that is developmentally appropriate for emerging readers and provides insight into how young children solve coding problems or use spatial knowledge to solve coding tasks. The purpose of the present study is to test the validity and reliability of a performance-based assessment that measures emerging readers' CT knowledge in the context of coding with tangible coding toys. The present study contributes to the research on early childhood CT, specifically, it contributes to understanding on how to assess CT as an active process (Bakala et al., 2021; Martins et al., 2023).

2.3. A note on assessment fairness

In the design and administration of the CaST assessment we are committed to fairness and ensuring that our inferences about children's CT learning are accurate (Clarke-Midura et al., 2021; Oliveri Elena et al., 2019). Similarly, we want to ensure that our items are not biased and that they do not favor children based on their gender or age (American Educational Research Association et al., 2014). Differential item functioning (DIF) analysis is one way to examine fairness in tests. DIF analysis allows us to determine if items are performing in a biased way towards members of a particular group. It is important to note that if boys score higher on an exam than girls, on average, it does not mean the items are biased in favor of boys. A DIF analysis is needed to determine if there is bias. While some of the research on assessing CT in early childhood reported difference in scores based on gender (Sullivan & Bers, 2013) and age (Zapata-Cáceres et al., 2020), they did not report results of DIF analysis. Thus, a contribution of the present study is that we conduct a DIF analysis to examine whether boys and girls and older and younger children, matched on ability, perform differently on any of the tasks in the assessment.

3. Methods

3.1. Participants

Our sample consisted of 272 children (girls = 138), between 47 and 101 months old ($M = 78.5$; $SD = 10.8$), across five elementary schools in the Western United States. For the analysis, age was split into two groups, younger (<72 months; $n = 95$) and older (≥72 months; $n = 177$). Selecting 72 months as a splitting point in the age is guided by the empirical work of prerequisite knowledge for the CT: math knowledge and spatial thinking (i.e., foundational ideas in our early childhood CT cognitive model). First, it is based on the validation studies on the *Research-based Early Math Assessment* (REMA, Alkhadim et al., 2021) that separate the sample at 72 months. Next, in the domain of spatial thinking, compared to younger children, 6-year-old children hold a comparable ability to adults in being aware of their mental rotation abilities and articulating them (Estes, 1998). Given that CaST is an interview-based performance assessment, awareness and explanation of their spatial activities can be critical criteria to determine a threshold to investigate developmental differences in the CT ability. In terms of prior experience with coding, 21 children reported doing coding activities at home and one school ($n = 123$) introduced coding activities starting in kindergarten.

3.2. Measures

We developed a standardized, interview-based assessment, *Computational and Spatial Thinking Assessment* (CaST), using the evidence-centered design framework (ECD). ECD is a systematic approach that involves constructing educational assessments in terms of evidentiary arguments (Mislevy & Haertel, 2006). This is done through an iterative process of observing what children say or do when completing tasks in order to make inferences about what they know and can do (Mislevy & Haertel, 2006). This view of assessment as argument is central to discussions around validity (American Educational Research Association et al., 2014; Kane et al., 2006) while offering what Mislevy (2007) calls "validity by design" where, as designers, we structure our approach in such a way that validity evidence emerges (Mislevy, 2007).

ECD consists of five layers: (a) domain analysis, (b) domain modeling, (c), the conceptual assessment framework, (d) assessment implementation, and (e) assessment delivery. In the first two layers, the focus is on the purposes of the assessment, the nature of knowing, and structures for observing and organizing knowledge. This information is put into "design patterns" that articulate the kinds of features that assessment tasks will need and the kinds of performances those features will elicit. In the third layer, the conceptual assessment framework (CAF), the focus is on the student model (what skills are being assessed), the evidence model (how do we measure it), and the task model (situations that elicit the behaviors/evidence). These three models are developed with the information from the first two layers in ECD – the design pattern in particular – to provide technical details of the tasks (such as potential student performances/products during assessment implementation and delivery) and a specification of the kinds of features of the tasks that will provide evidence about the student model. For example, we identified variable features, features that make a task vary in level of difficulty. These included administration features (enacting the program with agent and verbalizing codes); grid features (start space marked or unmarked, end space marked or unmarked); program features (program length, number of turns, location of turns), task features (starting orientation relative to students' perspective); and for items that involved debugging, the type and location of bug.

The assessment is unplugged and not tied to a specific coding platform. Children interact with 2-D grids and a tangible agent they can pick up and move along the grid (see Fig. 2). There are five different story lines involving a robot, each with a separate grid (or number line) and items that involve moving the robot from one location to another (e.g., putting a banana peel in the trash or picking up school supplies and putting them in a backpack). The tasks use four directional codes depicted by arrows to represent: *Forward, Backward, Rotate Right,* and *Rotate Left* (see Fig. 3). There are 19 task-based items that assess skills in our early childhood CT cognitive model (see Tables 1 and 2). Children responded to tasks by either ordering and sequencing codes, enacting a sequence of codes (by moving the agent), or debugging and fixing programs. Some tasks had only one possible correct solution whereas other tasks had multiple possible correct solutions. Each task was scored correct or incorrect. An example of an item is presented in Fig. 4. Note that the left presents the item from the perspective of the administration guide and the administrator's view of the grid while the image on the right presents what students see (the grid from a child's perspective). In this item, children are given a program that has an error. They are asked to "fix" the program by rewriting it. For more information about the assessment see (Clarke-Midura et al., 2021). Table 4 presents the CT concepts that are covered in each item.

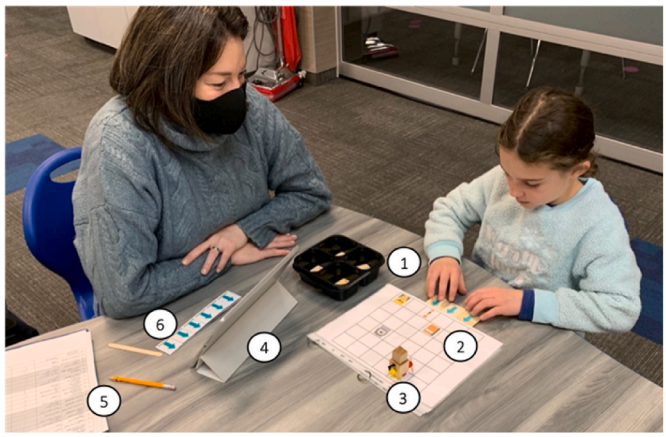


Fig. 2. Kindergarten Student Working on an Assessment Item Using (1) Arrow Codes, (2) Activity Grid, (3) Robot Agent, (4) Administrator's Assessment Scripts, (5) Scoring Sheets, and (6) Example Programs to Enact or Debug.

3.3. Procedure

The assessment was administered face-to-face in quiet areas in the schools, in a one-on-one format by members of the research team. All assessments were video recorded. Children were introduced to the assessment through a demonstration of how to use the four directional codes and through two sandbox items that were not scored. The assessments took an average 17.5 min ($SD = 3.3$) per child and were double scored based on video recordings.

3.4. Statistical analysis

Classical Test Theory (CTT) and Item Response Theory (IRT) are two common psychometric approaches to analyze and score test data (de Ayala, 2022). While they both provide useful information about test performance, we used IRT for the present study due to the ability to make stronger assumptions, such as the chance of getting items right or wrong based on a child's true ability. For example, when the assumptions of IRT are met, the parameters are sample and item independent, meaning a child will obtain the same true ability score (i.e., θ) no matter which a set of items within a given test that they answer (i.e., item independence) and items have the same difficulty and discrimination no matter which student is taking the test (i.e., sample independence).

Our analysis was conducted in five steps described below. We used the *ltm* package (Rizopoulos, 2006) for checking unidimensionality assumption, *mirt* package (Chalmersmirt, 2012) for IRT analyses, *difR* package (Magis et al., 2010) and *mirt* package (Chalmersmirt, 2012) for

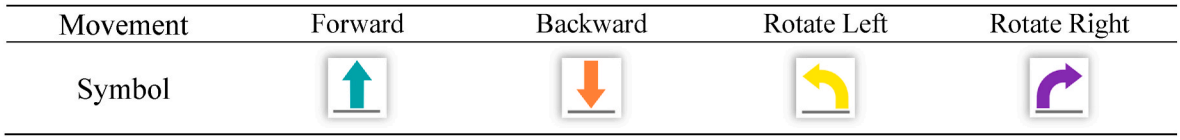


Fig. 3. Arrow Codes Used in the CaST Assessment.

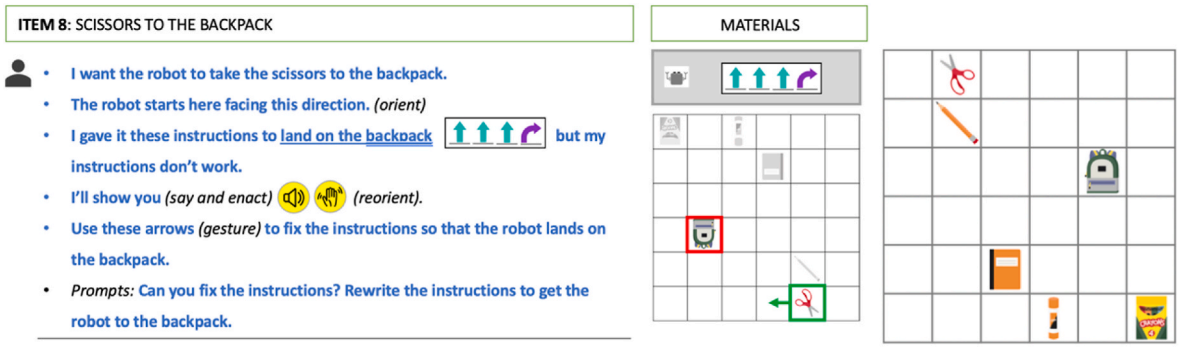


Fig. 4. An Example of CaST Assessment from the Administration Guide (Left) and Student Perspective (Right).

Table 4
CT Concepts Involved in Items.

Item	CT Concepts	Item	CT Concepts
Item 1	AT, Decomposition	Item 11	AT, Debugging
Item 2	AT	Item 12	AT
Item 3	AT	Item 13	AT, Debugging
Item 4	AT, Debugging	Item 14	AT, Decomposition
Item 5	AT	Item 15	AT, Decomposition
Item 6	AT, Debugging	Item 16	AT, Decomposition
Item 7	AT	Item 17	AT
Item 8	AT, Debugging	Item 18	AT
Item 9	AT	Item 19	AT, Debugging
Item 10	AT, Debugging		

Note. AT refers to algorithmic thinking.

DIF analyses, *psych* package (Revelle, 2023) for exploratory factor analyses (EFAs), *lavaan* package (Rosseel, 2012) for confirmatory factor analyses (CFAs), and *afex* package (Singmann et al., 2023) for ANOVAs. All analyses were conducted with R version 4.3.2 (R Core Team, 2023).

3.4.1. Assumptions check

To answer RQ 1, we employed IRT. We tested the three assumptions of IRT (see Table 5): (a) unidimensionality, (b) local independence, and (c) functional form (de Ayala, 2022).

3.4.1.1. Unidimensionality. For the assumption of unidimensionality, we performed exploratory factor analysis (EFA) and modified parallel analysis (Drasgow & Lissak, 1983) to check whether a single latent factor was held among students' item responses. In EFA, we extracted one factor and examined whether a single factor holds more than 20% of variance which is prerequisite to obtain stable parameter in the IRT framework (Reckase, 1979). Furthermore, in modified parallel analysis, satisfying the assumption of unidimensionality indicated that the second eigenvalues from the observed data was not substantially different from the second eigenvalues from the simulated data.

3.4.1.2. Local dependence. The second assumption is local dependence that the responses to one item should be independent of the responses to the other items. In the context of unidimensional scale, violating this assumption (i.e., local dependence) can lead to inflated reliability and weaken the accuracy in estimating person parameters. We used Yen's Q_3 statistics (Yen, 1984) to detect local dependence through the residual correlations among pairs of items. Considering both the sample size and the number of items (Christensen et al., 2017), we set a cutoff for the average residual correlation to lower than 0.3.

3.4.1.3. Functional form. The last assumption is the functional form that the given data should be fitted to the function specified by the model. The assumption was examined by fitting data to three different IRT models, including Rasch (variant of 1PL model), two (2PL) and three parameters (3PL) model. Among diverse model fit indices, Akaike information criterion (AIC, Akaike, 1974) and Bayes Information Criterion (BIC, Schwarz, 1978) were used to determine the final model. A model with the lowest AIC and BIC was selected as the best fitting model. As relative model fit, we also use loglikelihood and conducted model comparisons using log-likelihood ratio test.

3.4.2. Fitting IRT models

Next, we checked model level fit indices and item level fit indices. At the model level of goodness of fit, the limited-information statistics M_2 with its p-value, Root Mean Square Error of Approximation (RMSEA) and standardized root mean squared residual (SRMR) were comprehensively considered as indicators of the goodness of fit of the model

Table 5
Statistical Methods and Cutoff to Meet Three Assumptions of IRT.

Assumptions	Statistical Methods	Cutoff
1. Unidimensionality	Checking the proportion of variances explained by a single factor using exploratory factor analysis A comparison of observed and simulated second eigenvalues (Drasgow & Lissak, 1983)	Proportion of variance explained by a single factor >20% (Reckase, 1979)
2. Local Independence	Yen's Q_3 statistics (Yen, 1984)	$Q_3 < 0.3$ (Christensen et al., 2017)
3. Functional form	Model comparison to three IRT models through log-likelihood ratio test	

(Maydeu-Olivares & Joe, 2005, 2006). According to Maydeu-Olivares & Joe (2006), a non-significant M_2 with RMSEA below 0.089 and with SRMR below 0.05 were proposed cut-off values for good fit. At the item level of goodness of fit, we used signed chi-square ($S - \chi^2$) item-fit statistics (Orlando & Thissen, 2003) with RMSEA. The cutoff values for good fit were a non-significant $S - \chi^2$ with RMSEA close to 0.

3.4.3. Estimating item parameters and marginal reliability

After assessing model- and item-level fit, we fitted the selected model to the data to calibrate item parameters for the 19 items. For example, in the case of 2PL model, as a slope parameter, discrimination parameter (a) refers to how well items distinguish between the different levels of children's CT ability. The difficulty parameter (b) is a location parameter that reflects how difficult an item is. The 2PL IRT model can estimate person location (θ) based on their item response patterns and item parameters, which reflect the latent ability of children's CT. Additionally, we plotted item characteristic curves for the visual inspection of the relationships between item discrimination (a) and difficulty (b) according to children's CT latent ability (θ). We then examined test information function (TIF) to investigate the preciseness of test in the relationship to level of children's CT ability. The peak point of the test information function refers to where a test provides the most psychometric information, so it is the most reliable at measuring a child's CT ability. In addition to calculating the reliability of the composite test scores (i.e., Cronbach's α and McDonald's ω) under the framework of classical test theory, we estimated IRT marginal reliability (r_{xx}) (Cheng et al., 2012; Green et al., 1984) which is "the ratio of the true score variance to the total variance, expressed with respect to the estimated latent ability" (Andersson & Xin, 2018, p. 33). In the context of holding local independence assumption, the use of marginal reliability prevents overestimation of reliability and in turn allows for estimation of more precise reliability coefficients (Sireci et al., 1991).

3.4.4. Differential item functioning

We assessed differential item functioning (DIF) to check for biases in items. There are two widely used approaches for detecting DIF (Millsap, 2011): (a) observed variable analysis (e.g., sum score on a test) and (b) latent variable analysis (e.g., latent ability θ). The observed variable analysis posits this probability is dependent on the sum scores of test (s) which serves as a proxy to one's true ability (θ), whereas in latent variable analysis, the probability of responding correctly is conditioned on the latent ability (θ). Due to our relatively small sample size (Belzak, 2020) for conducting DIF analyses, we used a logistic regression with sum scores as an observed variable analysis (Swaminathan & Rogers, 1990) and two-parameter logistic IRT model with a log-likelihood test as a latent variable analysis (LRT) (Thissen et al., 1986) to detect potential DIF items. This approach first matches the sum scores (or latent ability (θ) in LRT) between two sub-groups and then statistically tests whether the relationships between a probability of correct answer and the total scores differ by two sub-groups.

Under the logistic regression with sum scores approach, we first built a baseline model for each item $_i$ (see Model 1) in which the probability of a correct item (p in Model 1) was regressed on the sum scores (s). Next, the sub-group variables (g) and the interaction between total scores and sub-group variables ($s \times g$) were entered into the baseline model (Model 2). When comparing the fit of Model 1 and Model 2, either $\beta_2 \neq 0$ (i.e., coefficient for the sub-group effect) or $\beta_3 \neq 0$ (i.e., coefficient for the interaction effect between sum scores and sub-group effect) indicate a DIF item.

$$\text{Model 1: } ln\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta_1 \times s.$$

$$\text{Model 2: } ln\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta_1 \times s + \beta_2 \times g + \beta_3 \times s \times g.$$

On the other hand, LRT compared the fit of two nested models: a compact model which assume that item parameters were equivalent between two groups, and an augmented model which allows for freely estimating item parameters (i.e., item difficulty and discrimination) for each group. To identify DIF items, LRT calculate the test statistics χ^2_C for a compact model and χ^2_A for an augmented model, respectively. If the differences in the test statistics between two models (i.e., χ^2_{LRT}) with two degrees of freedom (i.e., df_C - df_A) were statistically significant, it indicates the presence of a DIF effect in the item.

$$\chi^2_{LRT} = \chi^2_C - \chi^2_A$$

Using latent ability in detecting DIF allowed us to control for measurement error, and to detect intercept and slope DIF in the regression models (Belzak, 2020). While there are limitations of using the LRT with a small sample size, we conjecture that using both observed and latent variable analyses for DIF detection provides more robust evidence of DIF by comparing the two different results (Davidson et al., 2021). Further, to prevent inflation of type 1 error from the multiple comparisons, we adopted the Benjamini-Hochberg correction procedure (B–H) (Benjamini & Hochberg, 1995) to adjust alpha levels (Thissen et al., 2002).

3.4.5. Examining difference in latent ability of CT

We further examined the differences in children's CT ability by gender and age by conducting a two-way Analysis of Variance (ANOVA) with the latent ability (θ) of children's CT estimated from 2 PL IRT model as a dependent variable, and gender and age as factors.

4. Results

4.1. Descriptive statistics

Table 6 presents the proportion of correct item rates for all items by gender and age groups. Prior to the main analysis, we conducted a two-way ANOVA to identify (a) gender and age gaps in total CaST scores and (b) their potential interaction effects. There was no interaction effect between gender and age, $F(1, 268) = 0.59$, $p = 0.444$, $\eta_p^2 = 0.00$, whereas salient main effects were identified. Specifically, boys had higher mean scores of CaST than girls, $F(1, 268) = 5.75$, $p = 0.017$, $\eta_p^2 = 0.02$, and older children had higher mean scores of CaST than younger children, $F(1, 268) = 39.87$, $p < 0.001$, $\eta_p^2 = 0.13$.

4.2. Reliability

To check consistency in scoring process by two independent raters, Cohen's Kappa (Cohen, 1960), an index of inter-rater reliability, was calculated. The observed κ was 0.91, indicating a high level of inter-rater reliability (Landis & Koch, 1977). Internal reliability and general factor saturation were assessed for the 19 items in the form of Cronbach's alpha and McDonald's omega, respectively. The CaST assessment showed high internal consistency ($\alpha = 0.91$) and saturation ($\omega = 0.92$).

4.3. Verifying model assumptions

Table 5 summarizes statistical approaches and cutoffs to check three IRT assumptions. To check unidimensionality, we conducted EFA and confirmed that the first factor explained 31.0% of the variance, which was acceptable to meet unidimensionality (Reckase, 1979). The results of the modified parallel analysis (Drasgow & Lissak, 1983) showed that there was not a statistical difference ($p = 0.25$) between the second eigenvalues from the observed data ($\lambda = 1.33$) and simulated data ($\lambda = 1.22$) and a sharp elbow was detected between the number of first and second eigenvalue, both of which supported the unidimensionality (see Fig. 5).

For local independence, we calculated Yen's Q_3 statistics (Yen, 1984), the correlation between the residuals of pairs of items, and identified that there were two pairs of items with Q_3 statistic of 0.3 or

Table 6
Descriptive Statistics for CaST Correct Item Rates per Each Item by Gender and Age.

Item	Total (n = 272)		Gender				Age			
			Girl (n = 138)		Boy (n = 134)		Younger (n = 95)		Older (n = 177)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Item 1	0.64	0.48	0.55	0.50	0.73	0.45	0.51	0.50	0.71	0.45
Item 2	0.64	0.48	0.59	0.49	0.69	0.46	0.46	0.50	0.74	0.44
Item 3	0.50	0.50	0.44	0.50	0.57	0.50	0.34	0.48	0.59	0.49
Item 4	0.65	0.48	0.65	0.48	0.66	0.48	0.61	0.49	0.68	0.47
Item 5	0.40	0.49	0.37	0.48	0.43	0.50	0.30	0.46	0.45	0.50
Item 6	0.43	0.50	0.36	0.48	0.51	0.50	0.25	0.44	0.53	0.50
Item 7	0.22	0.41	0.20	0.40	0.23	0.42	0.11	0.31	0.28	0.45
Item 8	0.39	0.49	0.30	0.46	0.48	0.50	0.18	0.39	0.50	0.50
Item 9	0.42	0.49	0.36	0.48	0.49	0.50	0.24	0.43	0.51	0.50
Item 10	0.45	0.50	0.44	0.50	0.46	0.50	0.25	0.44	0.56	0.50
Item 11	0.43	0.50	0.37	0.48	0.49	0.50	0.26	0.44	0.51	0.50
Item 12	0.25	0.43	0.19	0.39	0.31	0.46	0.13	0.33	0.31	0.46
Item 13	0.34	0.47	0.31	0.47	0.36	0.48	0.19	0.39	0.41	0.49
Item 14	0.77	0.43	0.78	0.42	0.75	0.43	0.72	0.45	0.79	0.41
Item 15	0.90	0.30	0.88	0.32	0.92	0.28	0.83	0.38	0.94	0.24
Item 16	0.66	0.47	0.65	0.48	0.67	0.47	0.60	0.49	0.70	0.46
Item 17	0.79	0.41	0.79	0.41	0.78	0.41	0.64	0.48	0.86	0.34
Item 18	0.32	0.47	0.29	0.46	0.35	0.48	0.12	0.32	0.43	0.50
Item 19	0.64	0.48	0.67	0.47	0.60	0.49	0.50	0.50	0.71	0.45
Total	9.82	5.42	9.20	5.47	10.46	5.04	7.22	4.66	11.22	5.29

Note. M refers to mean, and *SD* refers to standardized deviation.

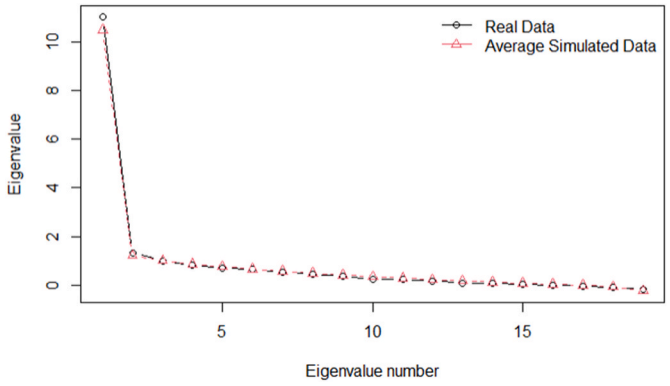


Fig. 5. Modified Parallel Analysis Plot for Undimensionality Assumption.

greater over than mean residual correlation, Items 2 and 3 ($r = 0.32$) and Items 9 and 12 ($r = -0.30$). However, eliminating the items yielded only minor changes in model- and item-fits. This indicated that including these items flagged in the local dependence is not problematic. For functional form, Table 7 presents model fits of three IRT models, Rasch, 2PL, and 3PL.The results of log-likelihood ratio test indicated that 2PL was a better fit than the Rasch model, $\chi^2(18) = 174.58$, $p < 0.001$; however, the 3PL model was not a better fit than the 2PL, $\chi^2(19) = 22.66$, $p = 0.253$. Therefore, we selected the 2PL model as the final model.

Table 7
Comparison of Three IRT Model Fit Indices (Rasch, 2PL and 3PL).

Model	AIC	BIC	LogLik	Model comparison	Loglikelihood Ratio Test
Rasch	4956.16	5028.28	−2458.08	–	–
2PL	4817.58	4954.60	−2370.79	Rasch vs. 2PL	$\chi^2(18) = 174.58$, $p < 0.001$
3PL	4832.92	5038.45	−2359.46	2PL vs. 3PL	$\chi^2(19) = 22.66$, $p = 0.253$

Note. AIC refers to AIC refers to Akaike information criterion; BIC refers to Bayesian information criterion; Loglik refers to log likelihood; LRT; 2PL refers to two-parameter item response theory model; 3PL refers to three-parameter item response theory model.

4.4. Assessing model and item-level fits

To assess model-level fit, we used M_2 with its accompanying p-value, *RMSEA*, and *SRMR*, jointly. Our results showed that, despite significant M_2 values ($M_2 = 248.00$, $p < 0.001$), other two metrics (*RMSEA* = 0.05, *SRMR* = 0.05) were good fits; thus, we regarded the results of model-level fit as acceptable. Furthermore, to assess item-level fit, we used the signed chi-square ($S - \chi^2$) item-fit statistics (Orlando & Thissen, 2003) and *RMSEA*. As a cutoff for the good model fit, $S - \chi^2$ is non-significant and *RMSEA* is recommended below 0.08. Table 8 shows all items fitted to the 2PL model.

4.5. Estimating item parameters

To interpret the estimated item parameters, we adopted Baker's guideline (Baker, 2001). According to the guideline (Baker, 2001), the theoretical range of item difficulty is from −4.0 to 4.0, but the practical

Table 8
Item-level Fit Indices.

Item	$S - \chi^2$			*RMSEA*
	$S - \chi^2$	*df*	*p*	
Item 1	22.68	14	0.07	0.05
Item 2	11.10	12	0.52	0.00
Item 3	7.32	12	0.84	0.00
Item 4	12.15	15	0.67	0.00
Item 5	9.29	12	0.68	0.00
Item 6	8.00	8	0.43	0.00
Item 7	8.74	8	0.37	0.02
Item 8	2.38	7	0.94	0.00
Item 9	11.31	8	0.19	0.04
Item 10	5.26	11	0.92	0.00
Item 11	10.75	10	0.38	0.02
Item 12	11.07	8	0.20	0.04
Item 13	13.72	12	0.32	0.02
Item 14	10.86	13	0.62	0.00
Item 15	7.39	9	0.60	0.00
Item 16	14.93	15	0.46	0.00
Item 17	5.71	9	0.77	0.00
Item 18	18.98	11	0.06	0.05
Item 19	9.80	13	0.71	0.00

Note. S − χ^2 refers to signed chi-square; *df* refers to degree of freedom; *RMSEA* refers to root-mean-square error of approximation.

range is −2.80 (very easy) to 2.80 (very difficult). Item discrimination was interpreted as "very low" for values between 0.01 and 0.34, "low" for values between 0.35 and 0.64, "moderate" for values between 0.65 and 1.34, "high" for values between 1.35 and 1.69, and "Very high" for values higher than 1.7 (Baker, 2001).

In Table 9, parameter estimates showed a wide range of item discrimination from moderate (Item 4 = 0.91) to very high (Item 6 = 4.66), with a high item discrimination across items, on average (M = 2.26, SD = 1.12). Item difficulty ranged from very easy (Item 15 = −2.40) to hard (Item 7 = 0.92), with a moderate level of difficulty across items ($M = -0.21$, $SD = 0.86$). Fig. 6 presents ICC for Item 15, which is on the far left, represents low difficulty and discrimination, whereas ICC for Item 7 and 12, which are on the far right, represent items with the highest item difficulties.

4.6. Test information function and marginal reliability

Fig. 7 shows the test information function (blue line in Fig. 7 left) with its standard errors (yellow dotted line in Fig. 7 left) and the marginal reliability according to the latent abilities of CT (Fig. 7 right). The test information function indicated that most item information is provided for children who possess average level of latent ability (θ, x-axis) from −1.0 to 1.5. The results implied that CaST assessment is the most precise measure for average CT ability (Max. item information ≅ 25.0; standard error of estimate ≅ 0.45), whereas this measure works poorly for children with extremely low or high latent abilities of CT. The marginal reliability (r_{xx}) for the CaST was 0.87, which is highly acceptable. Specifically, as shown in Fig. 7 right, the CaST assessment scores are reliable (i.e., $r_{xx} > 0.70$) in the medium range of the latent abilities of CT from (i.e., approximately $-1.5 < \theta < 1.8$ in x-axis), whereas this measure holds low reliability especially in the range of high latent abilities of CT (i.e., $\theta > 2.0$ in x-axis).

4.7. DIF analyses

To answer RQ 2, we conducted DIF analyses to check if the items functioned equally, regardless of a child's gender or age. Table 10 indicated that all items showed non-significant differences in B–H correction p-values by age or gender, indicating no item was flagged for either logistic regression with sum scores or LRT. This suggests that the 19 items measure children' CT latent ability fairly, regardless of their gender and age.

Table 9
Estimates of the Item Discrimination and Difficulty Parameters Using Two-parameter IRT Model.

	Item discrimination	Item difficulty
Item 1	1.16	−0.63
Item 2	1.69	−0.53
Item 3	2.16	−0.01
Item 4	0.91	−0.83
Item 5	2.18	0.34
Item 6	4.66	0.21
Item 7	3.01	0.92
Item 8	4.06	0.35
Item 9	3.65	0.25
Item 10	2.51	0.16
Item 11	3.12	0.23
Item 12	3.25	0.80
Item 13	2.27	0.55
Item 14	0.95	−1.47
Item 15	1.10	−2.40
Item 16	0.93	−0.86
Item 17	1.55	−1.19
Item 18	2.36	0.60
Item 19	1.46	−0.54

4.8. Differences in latent ability of CT by gender and age

As shown in Table 11, a two-way ANOVA result showed that the interaction effect was non-significant, $F(1, 268) = 0.76$, $p = 0.386$, $\eta_p^2 = 0.00$, but there were main effects of age, $F(1, 268) = 45.02$, $p < 0.001$, $\eta_p^2 = 0.14$, and gender, $F(1, 268) = 6.17$, $p = 0.014$, $\eta_p^2 = 0.02$. Specifically, boys ($M = 0.11$, $SD = 0.89$) showed significantly higher CT abilities than girls ($M = -0.11$, $SD = 0.98$). For age, the older group (≥72 months; $M = -0.26$, $SD = 0.90$) showed a higher latent ability of CT than the younger group (<72 months; $M = -0.48$, $SD = 0.84$) and its magnitude was large. Boys outperformed girls and older children outperformed younger children in the latent abilities of CT. Considering the results of both the ANOVA of latent ability and the DIF analyses, we can conclude that that the identified differences in CaST scores by gender and age (see 4.1. Descriptive statistics) were due to differences in true ability of CT, rather than due to items that were designed to favor certain subgroups.

5. Discussion

5.1. Discussion

5.1.1. Item characteristics of CaST

The goal of the present study is to provide evidence of validity, reliability, and fairness for CaST, a performance assessment of early childhood CT using IRT. Our results showed robust psychometric characteristics, including high reliability, unidimensionality, acceptable levels of item fit indices and parameters, and no evidence of DIF effects with a sample of 272 children ages 4–8.

IRT analyses showed that the assessment has good psychometric properties for young children. It has a high level of item discrimination ($M = 2.26$, $SD = 1.12$), on average, ranging from moderate to very high. Based on our test information function, the assessment provides the most information for children with average latent ability; it provides little information for children with low latent ability (i.e., below −1.5 SD below average) and high latent ability (i.e., above 2.0 SD above average) of CT. This result is different from previous findings on CT assessments for young children that found test information peaked at relatively low latent ability (El-Hamamsy et al., 2022; Relkin et al., 2020). In other words, TechCheck works well for discriminating between children with relatively low CT ability, whereas the assessment in the present study is better at reliably measuring children with average levels of CT. This suggests that our assessment has a distinct functionality to assess young children's CT that is not being met by other assessments. As mentioned above, early childhood researchers have not reached consensus on what skills constitute CT or how to assess CT. Our findings contribute to understanding about the kinds of CT abilities and skills that are developmentally appropriate for early childhood and how we can measure these skills in an active way, which we elaborate on in the next section. From a test development perspective, we now have evidence about how our items are performing that will help us re-design some of our items and design new items to specifically focus on children who fall in the lower and upper bands of CT ability.

5.1.2. Why certain items are more difficult for young children?

Similar to previous research on CT assessments for young children, we found that our assessment demonstrated moderate item difficulty levels ($M = -0.21$, $SD = 0.86$). However, as shown in Table 9, the item difficulty of our assessment varied widely. For example, looking at Table 9, item 15 has an item discrimination of 1.10 and item difficulty of −2.40. This was the easiest item on the assessment. Fig. 8 shows item 15 from the administration guide. In the item, children are asked to use a stick to mark where in the program the robot would stop at the tree. Some of the variable features in this item made it easy: the starting orientation of the robot is the same as the child's and the person administrating the assessment points to the two landmarks (the tree and

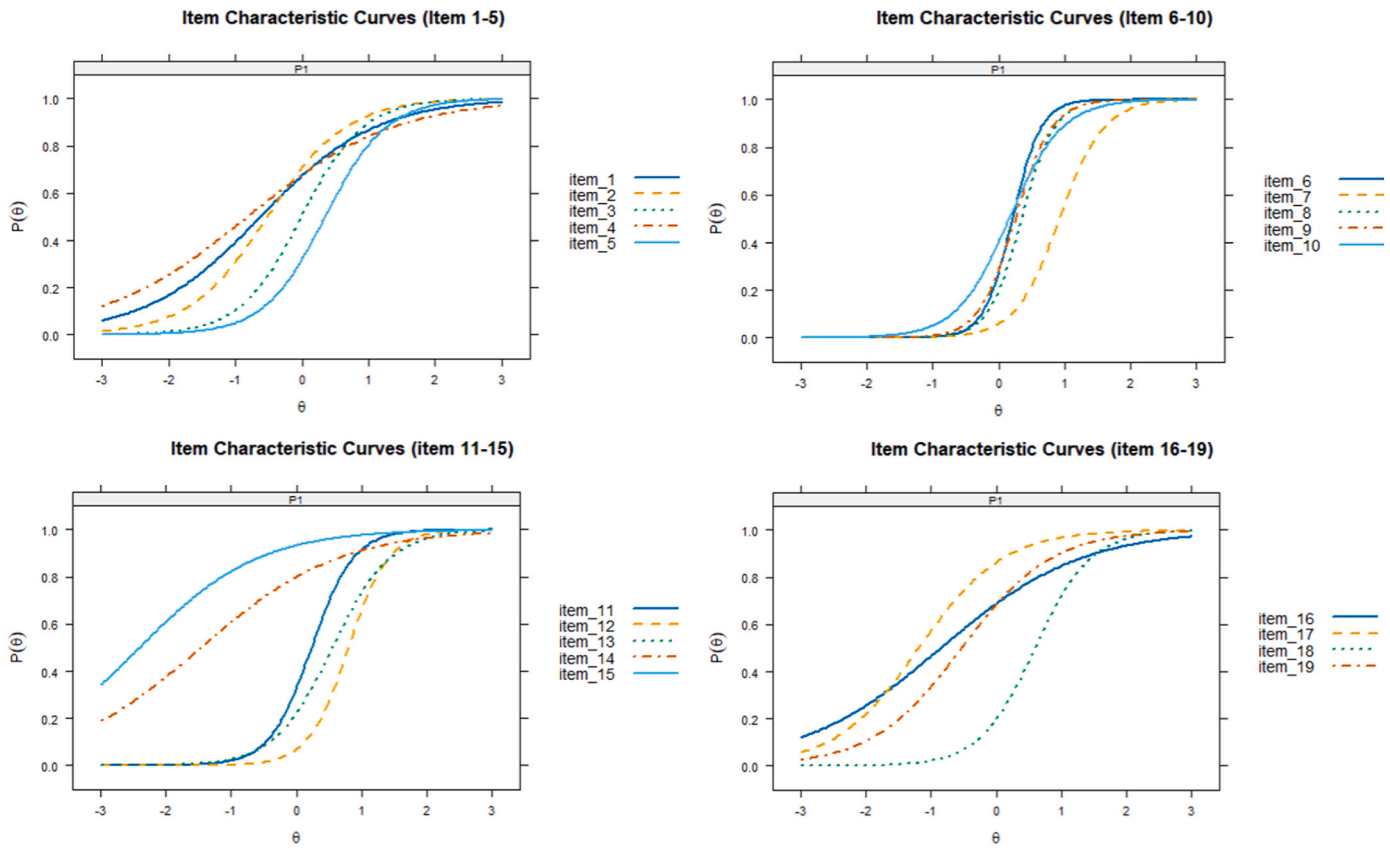


Fig. 6. Item Characteristic Curves for 19 Items of the Assessment.
Note. X-axis refers to children's latent ability (θ) of CT; Y-axis refers to the probability of the correct response.

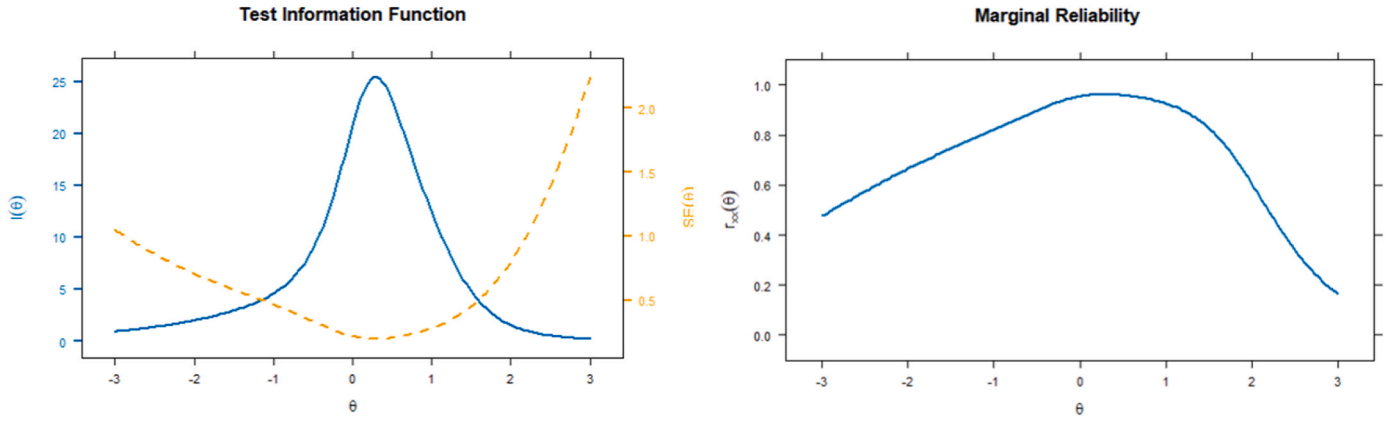


Fig. 7. Test Information Function with its Standard Errors (Left) and Marginal Reliability according to Different Levels of Latent Abilities (θ) of CT (Right).
Note. In the blue line of the left figure (Test Information Function), the X-axis refers to the children's latent ability (θ) of CT and the Y-axis refers to the amount of information provided by the item responses. In the yellow dotted line of the left figure, the X-axis refers to the children's latent ability (θ) of CT and the Y-axis refers to the standard error of the test information. In the right figure (Marginal Reliability), the X-axis refers to the children's latent ability (θ) of CT and the Y-axis refers to marginal reliability.

the house). This item assessed children's ability to coordinate the robot movement to the program and to decompose a program based on a landmark. This is an item that we will remove from our assessment because we want to reduce the number of items and this particular item does not provide a lot of discriminating information about children's understanding of CT.

As shown in Table 9, item 7 has a difficulty of 0.92, making it one of the hardest items for children in our sample. Fig. 9 depicts item 7 from the administrator's perspective where they read the script, orient the agent, and hand the child the program they are to enact. Fig. 10 shows the item from a child's perspective. It shows the item grid (A), the robot's starting orientation which is 90° to the left (B), and, if they move it correctly, the robot's ending location and orientation (C). This item assesses children's ability to enact a program when the robot does not share their orientation. Part of what makes item 7 difficult for children is that the robot's starting orientation is 90° to their left, which requires them to take on the perspective of the robot. It requires understanding that a forward movement is always a forward movement regardless of the robot's orientation. We found that many children enact the program from their own perspective or orientation (Jiang et al., 2023). Our

Table 10
Statistic of DIF Analysis for Gender and Age Using Logistic Regression with sum scores and 2-parameter Logistic IRT with Likelihood Ratio Tests.

	Gender						Age					
	Logistic regression with sum scores			LRT			Logistic regression with sum scores			LRT		
	Stat	*p*	*B–H*	$\Delta\chi^2$	*p*	*B–H*	*Stat*	*p*	*B–H*	$\Delta\chi^2$	*p*	*B–H*
Item 1	5.70	0.058	0.157	5.12	0.077	0.291	2.18	0.337	0.523	3.02	0.221	0.624
Item 2	3.69	0.158	0.308	1.75	0.417	0.495	2.12	0.346	0.523	2.21	0.331	0.648
Item 3	1.16	0.559	0.625	0.71	0.700	0.782	0.81	0.665	0.709	2.15	0.341	0.648
Item 4	2.04	0.360	0.456	3.29	0.193	0.334	2.76	0.252	0.523	2.72	0.257	0.624
Item 5	2.86	0.239	0.379	2.21	0.331	0.449	3.43	0.180	0.523	4.23	0.121	0.586
Item 6	3.64	0.162	0.308	6.14	0.046	0.291	4.89	0.087	0.413	4.37	0.112	0.586
Item 7	3.07	0.216	0.372	4.01	0.135	0.291	0.58	0.749	0.749	0.38	0.829	0.875
Item 8	5.81	0.055	0.157	5.41	0.067	0.291	2.69	0.261	0.523	1.71	0.425	0.674
Item 9	1.61	0.446	0.530	1.96	0.376	0.476	3.70	0.158	0.523	2.67	0.263	0.624
Item 10	6.79	0.034	0.157	8.03	0.018	0.171	2.06	0.358	0.523	1.01	0.605	0.751
Item 11	0.72	0.697	0.735	0.52	0.770	0.813	1.15	0.564	0.670	0.92	0.632	0.751
Item 12	4.59	0.101	0.239	3.49	0.174	0.331	0.80	0.672	0.709	0.69	0.707	0.790
Item 13	6.41	0.041	0.157	4.73	0.094	0.291	3.06	0.216	0.523	3.74	0.154	0.586
Item 14	2.59	0.274	0.400	2.26	0.323	0.449	1.34	0.513	0.670	1.40	0.498	0.675
Item 15	5.84	0.054	0.157	2.87	0.238	0.377	5.59	0.061	0.387	1.96	0.376	0.649
Item 16	2.23	0.328	0.445	3.96	0.138	0.291	6.69	0.035	0.387	4.19	0.123	0.586
Item 17	7.12	0.028	0.157	4.22	0.121	0.291	2.10	0.351	0.523	1.43	0.488	0.675
Item 18	0.39	0.824	0.824	0.37	0.830	0.830	6.24	0.044	0.387	5.22	0.074	0.586
Item 19	10.80	0.005	0.086	11.21	0.004	0.070	1.20	0.548	0.670	0.17	0.918	0.918

Note. LRT refers to 2-parameter logistic IRT model with log-likelihood tests; *B–H* refers to each *p*-value adjusted by Benjamini-Hochberg correction approach.

Table 11
The Result of a Two-way ANOVA by Age and Gender.

Factors	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*ES* (η_p^2)
Age	34.08	1	34.08	45.02	<0.001	0.14
Gender	4.67	1	4.67	6.17	0.014	0.02
Age × Gender	0.57	1	0.57	0.76	0.386	0.00
Error	202.91	268	0.76			

Note. df refers to degree of freedom; *F* refers to F-statistics; *ES* (η_p^2) refers to partial eta-partial squared as an effect size of measure.

findings are similar to studies on LOGO that found children tended to use their own perspective when programming the Turtle and struggled to program when they were not able to take on the Turtle's perspective

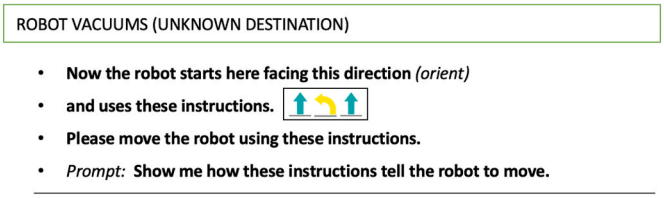


Fig. 9. Item 7 from Administration Guide Perspective.

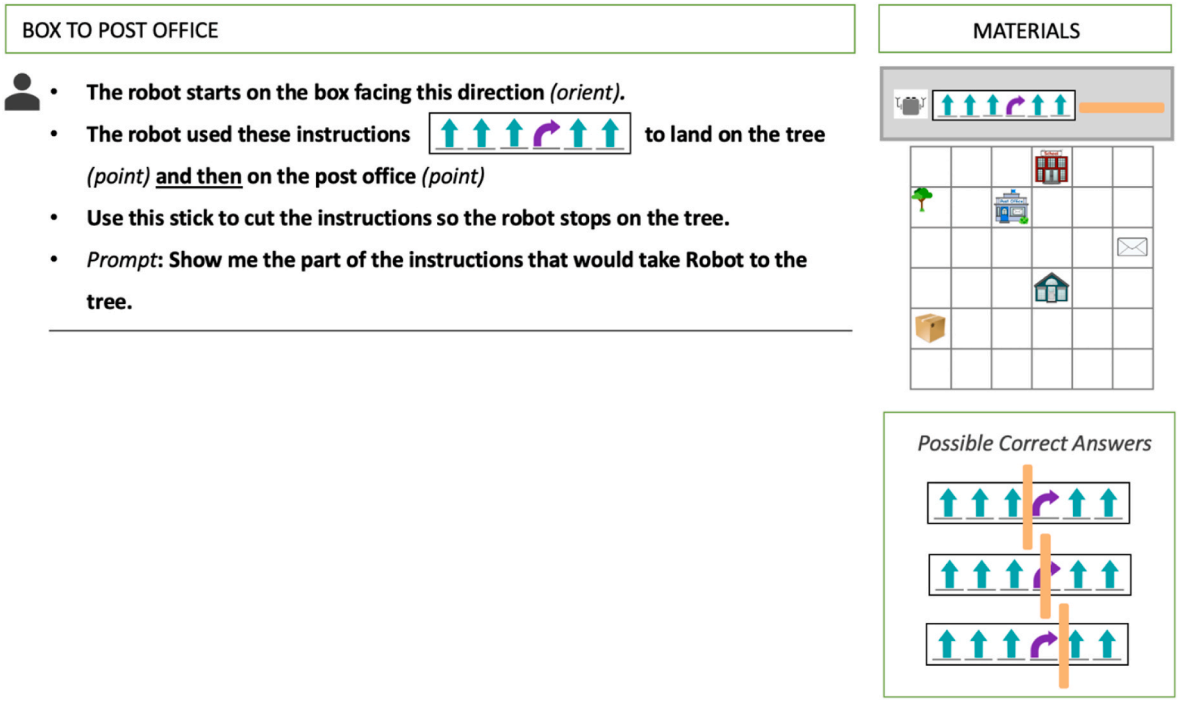


Fig. 8. Item 15 from Administration Guide Perspective.

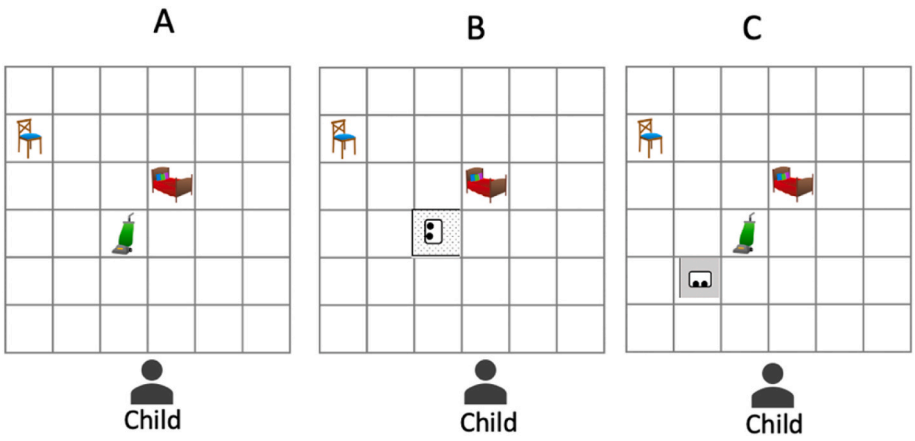


Fig. 10. Item 7 From Child's Perspective.
Note. A: Item grid for robot vacuums a room, B: Robot in starting orientation of 90° to the left of child's orientation, C: Robots ending location and orientation.

(Cuneo & Toronto, 1985; Fay & Mayer, 1987). For example, Mayer and Fay (1987) found that children often used an egocentric perspective when programming the Turtle in LOGO rather than a Turtle-centric perspective. The findings in the present study also align with recent studies that have shown relationships between concepts of spatial thinking and CT in early childhood (Città et al., 2019; Román-González et al., 2017; Tsarava et al., 2022).

As mentioned previously, most frameworks of CT do not contain or mention spatial thinking skills (Zeng et al., 2023). However, many environments that are used to teach coding in early childhood use tangible coding toys or coding that represents movement either in a physical or virtual space. Such movement requires that children understand spatial and mathematical concepts such as the codes always produce the same movements but depend on the agent's orientation, knowing that an agent's rotation occurs by rotating on a fixed point at a set angle, not translating to an adjacent point. The findings in the present study further contribute to a developmental understanding of CT for early childhood. Regardless of whether or not an item was easy or difficult, our findings provide useful information about what types of tasks are developmentally appropriate for young children or what skills are needed to engage in tasks designed around tangible coding toys or agents.

5.1.3. Test fairness and gender and age differences in CT ability

5.1.3.1. Test fairness through DIF. In terms of test fairness, we conducted a DIF analysis to explore whether our items favored children of a particular gender or age. We did not find any evidence of item bias based on gender and age. Our results are similar to the findings of de Ruiter and Bers (2022), who did not find any evidence of bias based on gender. However, they did find that one item showed evidence of DIF using Mentel-Haenszel tests based on age. Our findings indicate that our assessment can serve as an appropriate scale for comparing children's CT abilities by gender and age.

We also tested for main effects and potential interaction effects of age and gender on the children's latent ability (θ). While both main effects of gender and age were significant, their interaction effect was not significant. In terms of the non-significant interaction effect, this result is aligned with Sullivan and Bers (2016) in which no interaction effect between gender and grade levels was detected in robot and programming tasks. It may be due to the relatively narrow range of age range in the present study (4–8 years old), compared to previous studies (e.g., below 6 vs. 6–8 vs. above 8 years) (Rijke et al., 2018), which made it difficult to detect salient gender gaps. Future research is needed to recruit a wide range of children to explore whether the magnitude of gender differences in CT varies by age.

5.1.3.2. Differences in CT ability by age. We found that older children had significantly higher latent ability of CT than younger children. This finding is consistent with previous research that found children's age is a critical factor when designing and implementing lessons on CT and programming (Bati, 2022; McCormick & Hall, 2022). Saxena et al. (2020) designed CT activities with Bee-Bot to teach algorithm design for two age groups: K-1 (aged 3 to 4) and K-2 (aged 5 to 6). While K-2 children mastered algorithm design, K-1 children only partially solved the tasks and struggled with directional language. These findings highlight the need for developmentally appropriate practices in early childhood (NAEYC & Fred Rogers Center for Early Learning and Children's Media, 2012). Technology and media have potential to enhance children's learning experiences when educators make decisions carefully and the integration of these media into learning activities are developmentally appropriate for children. Similarly, researchers have been calling for developmentally appropriate assessments of CT for early childhood (Clarke-Midura et al., 2023; Relkin et al., 2023). Relkin et al. (2023) recently proposed a grade-specific CT assessment, TechCheck K, 1, and 2 and its normalization scoring system. While the work of Relkin et al. and our findings contribute to understanding of how to assess CT in early childhood, there is a need for further research on what it means to design and implement developmentally appropriate CT curricula and assessments in early childhood classrooms.

5.1.3.3. Differences in CT ability by gender. The present study found that boys outperformed girls on the CaST assessment and that this difference was not due to bias in the items but due to children's true latent CT ability. We offer three possible explanations for these differences: lack of developmentally appropriate and meaningful curriculum, the use of sum scores, and spatial thinking, which we discuss briefly below.

Findings on gender and CT ability in early childhood are mixed with some studies indicating that CT abilities were not predicted by a child's gender (El-Hamamsy et al., 2022; Papadakis et al., 2016; Relkin et al., 2020) and other studies suggesting boys had higher CT abilities (Angeli & Valanides, 2020; Sullivan & Bers, 2013). In a recent systematic review on CT and programming in early childhood, Bati (2022) reported that studies on early childhood CT were more likely to find differences in gender related to motivational and social factors rather than ability or performance. They suggested that a potential reason for any gender differences is due to the lack of developmentally and suitable content related to children's needs. For example, Sullivan and Bers (2013) found that girls and boys perform similarly on items related to concepts such as debugging but boys score higher on conditionals and fitting robot gear. Angeli and Valanides (2020) found that boys benefit more from individual, kinesthetic, spatially directed, and manipulative-based activities with cards and girls benefit more from collaborative writing activities

(Bati, 2022). While most of the research on broadening participation of women in CS is focused on middle school through college, findings suggest that CS education often emphasizes curriculum, tools, and materials that are historically aligned more closely with male interests than female interests (Peppler & Wohlwend, 2018) and females prefer collaboration, real-world projects, and those that emphasize creativity and aesthetics (Buechley & Hill, 2010; Guzdial et al., 2012; Margolis et al., 2011). This further supports the need for developmentally appropriate curriculum related to children's interests and needs in early childhood settings.

Another possible explanation for why our findings tell a slightly different story on gender differences in CT ability than previous studies is because most of the other studies used the sum of observed scores whereas the present study used latent ability scores (θ), which control for the measurement errors as a proxy to children's CT ability. These different approaches to scoring (i.e., whether measurement errors are controlled or not) have potential to result in inconsistent patterns of gender difference in CT abilities.

Finally, the findings on gender differences in CT ability could be related to differences in spatial thinking ability. CaST is different from other CT measures due to its focus on the spatial thinking that is required to solve some of the tasks. Research has found that boys perform better on tests of spatial thinking than girls. For example, a recent meta-analysis on spatial thinking (Lauer et al., 2019) found that gender differences in spatial ability occurred from an early age ($g = 0.20$ for 3–7 years). As mentioned above, as far back as LOGO, researchers have been documenting the relationship between spatial thinking and CT. It could be that boys in our sample started kindergarten with more spatial thinking knowledge than the girls, which influenced their performance on CaST. More research is needed to understand the relationship between spatial thinking and CT. Overall, the findings in the present study indicate the importance of finding developmentally appropriate ways to use tangible coding toys in preschool and kindergarten classrooms in order to provide girls with experiences and opportunities to play with tangible coding toys in meaningful ways.

5.2. Limitations and suggestions for future research

Several limitations of the current study should be considered when interpreting the potential strengths and psychometric evidence of the CaST. First, our results showed that 19 CaST items are well fitted to 2PL IRT model, test information function (Fig. 7 left) showed that most item information is centered around the range of −1.0 to 1.5 of the X-axis (θ, latent abilities of CT). This indicates that CaST functions well to assess average ability of children, whereas this measure is less sensitive to assess children with very low or high levels of CT. In future work we plan to develop more items and modify some of the variable features to extend the range of our test information.

Second, the sample size in the present study might be regarded as modest. However, according to de Ayala (2022), determining the sample size for IRT analysis should involve multiple considerations, including the type of response, the length of the item, person distribution and the number of parameters to be estimated, rather than hard-and-fast rules. Further, Morizot et al. (2007) stated that for dichotomously scored items, it is possible to have as few as 200 participants for unbiased analyses for 2PL IRT models. Accordingly, the sample size of 272 in the present study is sufficient to fit the IRT model with dichotomous responses. However, administering CaST with a large sample in a future study will allow for stable and accurate parameter estimations under more complex IRT models.

Third, our efforts to validate CaST substantially are aligned with sources for validity evidence: content, cognitive process, internal structure (see Appendix A), conceptually related constructs, and consequence of testing (American Educational Research Association et al., 2014). However, we could not confirm evidence on the *relationships with criteria* due to the restrictions on administration of other assessments. Future studies should investigate not only the level of association with other early childhood CT assessments, but also the predictive relationships with general cognitive abilities.

Lastly, as a potential moderator of explaining differences in CT abilities, the current study focused on investigation of gender and age. Although potential gender gaps and developmental appropriateness of assessment of early childhood CT are critical to consider, other demographic factors should also be included as potential moderators of CT ability in future studies.

6. Conclusion

Early childhood computational thinking is an emerging field (Zeng et al., 2023). The present study contributes to knowledge and understanding of early childhood CT by providing evidence for using performance-based assessments to make valid and reliable inferences about young children's CT knowledge that is not dependent on their reading or writing ability. Our findings show that gender differences in CT understanding exists as early as primary school and indicate the importance of providing girls opportunities to play with tangible coding toys and to gain foundational spatial, math, and CT skills early in preschool and primary school. Finally, while children's access to tangible coding toys and apps has increased, more research is needed on how we can support teachers' use of these tools for meaningful CT learning in early childhood classrooms.

7. Selection and participation

Data for this study were collected at five different schools in the Rocky Mountain region of the U.S. All children participated in this study voluntarily. In accordance with the Institutional Review Broad protocols, we followed a two-step process in which parents were first invited to consent to their child's participation. Children who had parental consent were then invited to participate, were informed about the data collection procedures, and told that they could opt out or withdraw at any time. Children then provided verbal assent to participate in the study.

Funding

This work was supported in part by funding from the National Science Foundation under Grant No. DRL- 1842116.

CRediT authorship contribution statement

Chungsoo Na: Data curation, Formal analysis, Writing – original draft. **Jody Clarke-Midura:** Conceptualization, Funding acquisition, Investigation, Project administration, Supervision, Writing – original draft. **Jessica Shumway:** Funding acquisition, Investigation, Writing – review & editing. **Wilhelmina van Dijk:** Validation, Writing – review & editing. **Victor R. Lee:** Funding acquisition, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Data availability

The authors do not have permission to share data.

Appendix A. Internal Structure of CaST Assessment

To find validity evidence of internal structure, we conducted a confirmatory factor analysis (CFA) with diagonally weighted least squares (DWLS) estimator, and the results showed acceptable model fit indices, $\chi^2(152) = 242.20$, $p < 0.001$, CFI = 0.99, TLI = 0.99, RMSEA = 0.05, 90% CI [0.04, 0.06], SRMR = 0.08. The overall factor loadings were higher than 0.4 across 19 items. The CFA results support validity evidence of internal structure and unidimensionality of the assessment.

Table A1
The Results of Confirmatory Factor Analysis.

	Factor Loading	*SE*	*z-score*	Standardized Factor Loading	*p-value*
Item 1	1.00			0.59	<0.001
Item 2	1.25	0.15	8.14	0.74	<0.001
Item 3	1.40	0.15	9.18	0.83	<0.001
Item 4	0.86	0.14	6.32	0.51	<0.001
Item 5	1.34	0.16	8.65	0.79	<0.001
Item 6	1.61	0.17	9.51	0.96	<0.001
Item 7	1.42	0.15	9.26	0.84	<0.001
Item 8	1.57	0.17	9.39	0.93	<0.001
Item 9	1.55	0.17	9.38	0.92	<0.001
Item 10	1.42	0.16	8.74	0.84	<0.001
Item 11	1.51	0.16	9.24	0.89	<0.001
Item 12	1.48	0.16	9.36	0.88	<0.001
Item 13	1.36	0.15	8.81	0.80	<0.001
Item 14	0.81	0.15	5.31	0.48	<0.001
Item 15	0.86	0.18	4.87	0.51	<0.001
Item 16	0.86	0.14	5.95	0.51	<0.001
Item 17	1.13	0.16	7.18	0.67	<0.001
Item 18	1.37	0.15	9.05	0.81	<0.001
Item 19	1.13	0.14	8.01	0.67	<0.001

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723. https://doi.org/10.1109/TAC.1974.1100705

Alkhadim, G. S., Cimetta, A. D., Marx, R. W., Cutshaw, C. A., & Yaden, D. B. (2021). Validating the research-based early math assessment (REMA) among rural children in southwest United States. *Studies In Educational Evaluation, 68*, Article 100944. https://doi.org/10.1016/j.stueduc.2020.100944

American Educational Research Association. (2014). *American psychological association, national council on measurement in education, & joint committee on standards for educational and psychological testing (U.S.), standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andersson, B., & Xin, T. (2018). Large sample confidence intervals for item response theory reliability coefficients. *Educational and Psychological Measurement, 78*, 32–45. https://doi.org/10.1177/0013164417713570

Angeli, C., & Valanides, N. (2020). Developing young children's computational thinking with educational robotics: An interaction effect between gender and scaffolding strategy. *Computers in Human Behavior, 105*, Article 105954. https://doi.org/10.1016/j.chb.2019.03.018

Bakala, E., Gerosa, A., Hourcade, J. P., & Tejera, G. (2021). Preschool children, robots, and computational thinking: A systematic review. *Int. J. Child-Comput. Interact., 29*, Article 100337. https://doi.org/10.1016/j.ijcci.2021.100337

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.) Retrieved from https://eric.ed.gov/?id=ED458219.

Bati, K. (2022). A systematic literature review regarding computational thinking and programming in early childhood education. *Education and Information Technologies, 27*, 2059–2082. https://doi.org/10.1007/s10639-021-10700-2

Belzak, W. C. M. (2020). Testing differential item functioning in small samples. *Multivariate Behavioral Research, 55*, 722–747. https://doi.org/10.1080/00273171.2019.1671162

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol., 57*, 289–300.

Bers, M. U. (2018). *Coding as a playground: Programming and computational thinking in the early childhood classroom* (1st ed.). New York: Routledge.

Brennan, K., & Resnick, M. (2012). New frameworks for studying and assessing the development of computational thinking. In *Proc. 2012 annu. Meet. Am. Educ. Res. Assoc. Vanc. Can* (p. 25).

Buechley, L., & Hill, B. M. (2010). LilyPad in the wild: How hardware's long tail is supporting new engineering and design communities. In *Proc. 8th ACM conf. Des. Interact. Syst.* (pp. 199–207). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1858171.1858206.

Chalmers, R. P., & mirt. (2012). A Multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*, 1–29. https://doi.org/10.18637/jss.v048.i06

Cheng, Y., Yuan, K.-H., & Liu, C. (2012). Comparison of reliability measures under factor analysis and item response theory. *Educational and Psychological Measurement, 72*, 52–67. https://doi.org/10.1177/0013164411407315

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement, 41*, 178–194. https://doi.org/10.1177/0146621616677520

Città, G., Gentile, M., Allegra, M., Arrigo, M., Conti, D., Ottaviano, S., Reale, F., & Sciortino, M. (2019). The effects of mental rotation on computational thinking. *Computer Education, 141*, Article 103613. https://doi.org/10.1016/j.compedu.2019.103613

Clarke-Midura, J., Lee, V. R., Shumway, J. F., Silvis, D., Kozlowski, J. S., & Peterson, R. (2023). Designing formative assessments of early childhood computational thinking. *Early Childhood Research Quarterly, 65*, 68–80.

Clarke-Midura, J., Silvis, D., Shumway, J. F., Lee, V. R., & Kozlowski, J. S. (2021). Developing a kindergarten computational thinking assessment using evidence-centered design: the case of algorithmic thinking. *Computer Science Education, 31*(2), 117–140.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46. https://doi.org/10.1177/001316446002000104

Cuneo, D. O. (1985). In O. Toronto (Ed.), *Young children and turtle graphics programming: Understanding turtle commands*. Canada. Retrieved from https://eric.ed.gov/?id=ED260800. (Accessed 6 May 2023).

Davidson, M. J., Wortzman, B., Ko, A. J., & Li, M. (2021). Investigating item bias in a CS1 exam with differential item functioning. In *Proc. 52nd ACM tech. Symp. Comput. Sci. Educ.* (pp. 1142–1148). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3408877.3432397.

de Ayala, R. J. (2022). *The theory and practice of item response theory* (2nd ed.). Guilford Publications.

de Ruiter, L. E., & Bers, M. U. (2022). The coding stages assessment: Development and validation of an instrument for assessing young children's proficiency in the ScratchJr programming language. *Computer Science Education, 32*, 388–417. https://doi.org/10.1080/08993408.2021.1956216

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68*, 363–373. https://doi.org/10.1037/0021-9010.68.3.363

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

Estes, D. (1998). Young children's awareness of their mental activity: The case of mental rotation. *Child Development, 69*, 1345–1360. https://doi.org/10.2307/1132270

Fay, A. L., & Mayer, R. E. (1987). Children's naive conceptions and confusions about Logo graphics commands. *Journal of Educational Psychology, 79*, 254–268. https://doi.org/10.1037/0022-0663.79.3.254

Govind, M., & Bers, M. (2021). Assessing robotics skills in early childhood: Development and testing of a tool for evaluating children's projects. *J. Res. STEM Educ., 7*, 47–68. https://doi.org/10.51355/jstem.2021.102

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347–360. https://doi.org/10.1111/j.1745-3984.1984.tb01039.x

Guzdial, M., Ericson, B. J., McKlin, T., & Engelman, S. (2012). A statewide survey on computing education pathways and influences: Factors in broadening participation in computing. In *Proc. Ninth annu. Int. Conf. Int. Comput. Educ. Res.* (pp. 143–150). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2361276.2361304.

Heikkilä, M., & Mannila, L. (2018). Debugging in programming as a multimodal practice in early childhood education settings. *Multimodal Technol. Interact., 2*. https://doi.org/10.3390/mti2030042

Howard, S. J., & Melhuish, E. (2017). An early years toolbox for assessing early executive function, language, self-regulation, and social development: Validity, reliability, and preliminary norms. *Journal of Psychoeducational Assessment, 35*, 255–275. https://doi.org/10.1177/0734282916633009

Jiang, M., Clarke-Midura, J., Silvis, D., Shumway, J., & Lee, V. R. (2023). Which Way is Up? Orientation and Young Children's Directional Arrow Interpretations in Coding Contexts. *International Society of the Learning Sciences*.

Kane, M. (2006). In S. M. Downing, & T. M. Haladyna (Eds.), *Handb. Test dev.Content-related validity evidence in test development* (pp. 131–153). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. https://doi.org/10.2307/2529310

Lauer, J. E., Yhang, E., & Lourenco, S. F. (2019). The development of gender differences in spatial reasoning: A meta-analytic review. *Psychological Bulletin, 145*, 537–565. https://doi.org/10.1037/bul0000191

Macrides, E., Miliou, O., & Angeli, C. (2022). Programming in early childhood education: A systematic review. *Int. J. Child-Comput. Interact., 32*, Article 100396. https://doi.org/10.1016/j.ijcci.2021.100396

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*, 847–862. https://doi.org/10.3758/BRM.42.3.847

Margolis, J., Goode, J., & Bernier, D. (2011). The need for computer science. *Education Leader, 68*, 68–72.

Martins, E. C., da Silva, L. G. Z., & Neris, V. P. de A. (2023). Systematic mapping of computational thinking in preschool children. *Int. J. Child-Comput. Interact., 36*, Article 100566. https://doi.org/10.1016/j.ijcci.2023.100566

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2n contingency tables. *Journal of the American Statistical Association, 100*, 1009–1020. https://doi.org/10.1198/016214504000002069

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*, 713–732. https://doi.org/10.1007/s11336-005-1295-9

Mayer, R. E., & Fay, A. L. (1987). A chain of cognitive changes with learning to program in Logo. *Journal of Educational Psychology, 79*, 269–279. https://doi.org/10.1037/0022-0663.79.3.269

McCormick, K. I., & Hall, J. A. (2022). Computational thinking learning experiences, outcomes, and research in preschool settings: A scoping review of literature. *Education and Information Technologies, 27*, 3777–3812. https://doi.org/10.1007/s10639-021-10765-z

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY, US: Routledge/Taylor & Francis Group.

Mislevy, R. J. (2007). Validity by design. *Educational Research, 36*, 463–469. https://doi.org/10.3102/0013189X07311660

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*, 6–20. https://doi.org/10.1111/j.1745-3992.2006.00075.x

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handb. Res. Methods personal. Psychol.* (pp. 407–423). New York, NY, US: The Guilford Press.

NAEYC & Fred Rogers Center for Early Learning and Children's Media. (2012). *Technology and interactive media as tools in early childhood programs serving children from birth through age 8 (Joint position statement), Latrobe, PA*. Retrieved from https://www.naeyc.org/sites/default/files/globally-shared/downloads/PDFs/resources/position-statements/ps_technology.pdf.

Oliveri Elena, M., Lawless, R., & Mislevy, R. J. (2019). Using evidence-centered design to support the development of culturally and linguistically sensitive collaborative problem-solving assessments. *International Journal of Testing, 19*, 270–300. https://doi.org/10.1080/15305058.2018.1543308

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit Index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*, 289–298. https://doi.org/10.1177/0146621603027004004

Papadakis, S., Kalogiannakis, M., & Zaranis, N. (2016). Developing fundamental programming concepts and computational thinking with scratchJr in preschool education: A case study. *International Journal of Mobile Learning and Organisation, 10*, 187–202. https://doi.org/10.1504/IJMLO.2016.077867

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

Peppler, K., & Wohlwend, K. (2018). Theorizing the nexus of STEAM practice. *Arts Education Policy Review, 119*, 88–99. https://doi.org/10.1080/10632913.2017.1316331

R Core Team. (2023). *R: A language and environment for statistical computing*. https://www.R-project.org/.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207–230. https://doi.org/10.2307/1164671

Relkin, E., & Bers, M. U. (2019). Designing an assessment of computational thinking abilities for young children. In L. E. Cohen, & S. Waite-Stupiansky (Eds.), *STEM early child. Learn. Sci. Technol. Eng. Math. Strengthen learn.* (pp. 83–98). Routledge.

Relkin, E., & Bers, M. (2021). TechCheck-K: A measure of computational thinking for kindergarten children. In *2021 IEEE glob. Eng. Educ. Conf. EDUCON* (pp. 1696–1702). https://doi.org/10.1109/EDUCON46332.2021.9453926

Relkin, E., de Ruiter, L., & Bers, M. U. (2020). TechCheck: Development and validation of an unplugged assessment of computational thinking in early childhood education. *Journal of Science Education and Technology, 29*, 482–498. https://doi.org/10.1007/s10956-020-09831-x

Relkin, E., Johnson, S. K., & Bers, M. U. (2023). A normative analysis of the TechCheck computational thinking assessment. *Educational Technology & Society, 26*, 118–130. https://doi.org/10.30191/ETS.202304_26(2).0009

Revelle, W. (2023). *Procedures for psychological, psychometric, and personality research*. psych: https://CRAN.R-project.org/package=psych.

Rijke, W. J., Bollen, L., Eysink, T. H. S., & Tolboom, J. L. J. (2018). Computational thinking in primary school: An examination of abstraction and decomposition in different age groups. *Informatics in Education, 17*, 77–92. https://doi.org/10.15388/infedu.2018.05

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response analysis. *Journal of Statistical Software, 17*, 1–25. https://doi.org/10.18637/jss.v017.i05

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

Rosseel, Y. (2012). lavaan: An R Package for structural equation modeling. *Journal of Statistical Software, 48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Saxena, A., Lo, C. K., Hew, K. F., & Wong, G. K. W. (2020). Designing unplugged and plugged activities to cultivate computational thinking: An exploratory study in early childhood education. *Asia-Pac. Educ. Res., 29*, 55–66. https://doi.org/10.1007/s40299-019-00478-w

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Shumway, J. F., Clarke-Midura, J., Lee, V. R., Silvis, D., Welch Bond, & Kozlowski, J. S. (2023). *Teaching Coding in Kindergarten: Supporting Students' Activity with Robot Coding Toys. In Teaching Coding in K-12 Schools: Research and Application* (pp. 23–38). Cham: Springer International Publishing.

Shumway, J. F., Welch, L. E., Kozlowski, J. S., Clarke-Midura, J., & Lee, V. R. (2023). Kindergarten students' mathematics knowledge at work: the mathematics for programming robot toys. *Mathematical Thinking and Learning, 25*(4), 380–408.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2023). *afex: Analysis of factorial experiments*. https://CRAN.R-project.org/package=afex.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237–247. https://doi.org/10.1111/j.1745-3984.1991.tb00356.x

Su, J., & Yang, W. (2023). A systematic review of integrating computational thinking in early childhood education. *Comput. Educ. Open, 4*, Article 100122. https://doi.org/10.1016/j.caeo.2023.100122

Sullivan, A., & Bers, M. U. (2013). Gender differences in kindergarteners' robotics and programming achievement. *International Journal of Technology and Design Education, 23*, 691–702. https://doi.org/10.1007/s10798-012-9210-z

Sullivan, A., & Bers, M. U. (2016). Girls, boys, and bots: Gender differences in young children's performance on robotics and programming tasks. *Journal of Information Technology Education: Innovations in Practice, 15*, 145–165.

Sung, J. (2022). Assessing young Korean children's computational thinking: A validation study of two measurements. *Education and Information Technologies, 27*, 12969–12997. https://doi.org/10.1007/s10639-022-11137-x

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

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

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118–128. https://doi.org/10.1037/0033-2909.99.1.118

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27*, 77–83. https://doi.org/10.3102/10769986027001077

Tsarava, K., Moeller, K., Román-González, M., Golle, J., Leifheit, L., Butz, M. V., & Ninaus, M. (2022). A cognitive definition of computational thinking in primary education. *Computer Education, 179*, Article 104425. https://doi.org/10.1016/j.compedu.2021.104425

Unahalekhaka, A., & Bers, M. U. (2022). Evaluating young children's creative coding: Rubric development and testing for ScratchJr projects. *Education and Information Technologies, 27*, 6577–6597. https://doi.org/10.1007/s10639-021-10873-w

Examining computational thinking in early childhood [special issue], early childWang, X. C., Bers, M. U., & Lee, V. R. (Eds.). *Restoration Quarterly, 65*, (2023), 42–158.

Wang, X. C., Flood, V. J., & Cady, A. (2021). Computational thinking through body and ego syntonicity: Young children's embodied sense-making using a programming toy. In E. de Vries, Y. Hod, & J. Ahn (Eds.), *Proc. 15th int. Conf. Learn. Sci. - icls 2021* (pp. 394–401). Bochum, Germany: International Society of the Learning Sciences. https://doi.org/10.22318/icls2021.394.

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

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–145. https://doi.org/10.1177/014662168400800201

Zapata-Cáceres, M., Martín-Barroso, E., & Román-González, M. (2020). Computational thinking test for beginners: Design and content validation. In *2020 IEEE glob. Eng. Educ. Conf. EDUCON* (pp. 1905–1914). https://doi.org/10.1109/EDUCON45650.2020.9125368

Zeng, Y., Yang, W., & Bautista, A. (2023). Computational thinking in early childhood education: Reviewing the literature and redeveloping the three-dimensional framework. *Educational Research Review, 39*, Article 100520. https://doi.org/10.1016/j.edurev.2023.100520