International Society of the Learning Sciences

Assessing Young Children's Computational Thinking Using Cognitive Diagnostic Modeling

Chungsoo Na, Jody Clarke-Midura
chungsoo.na@outlook.com, jody.clarke@usu.edu
Utah State University

Abstract: This study illustrates how Cognitive Diagnostic Modeling (CDM) can be used to assess fine-grained levels of computational thinking (CT). We analyzed scored responses to the *Computational and Spatial Thinking assessment* (CaST) from 271 children. We identified four key attributes required to solve tasks: *sequencing of codes, fixing a program, spatial orientation of an agent, and rotation on a point*. Results indicated that younger children did not master all the attributes, particularly *spatial orientation of an agent* and *rotation on a point*. We identified four common mastery profiles of children that were associated with age. Our findings illustrate that mastering *spatial orientation* is critical to CT ability. Finally, the nuanced information about children's mastery levels has potential to provide teachers with useful information about what concepts and skills their students are struggling with so that they can adjust instruction to emphasize those concepts.

Introduction

Providing children opportunities to engage in computational thinking (CT) in early elementary school is becoming increasingly important. CT is "the thought processes involved in formulating a problem and expressing its solution" (Wing, 2014) and is often operationalized in the context of coding. Accordingly, there is a need for a better understanding of the instructional approaches, practices, and assessments that support children's development of CT in early elementary classrooms (Luo et al., 2022).

While instructional resources and assessments of CT for early elementary exist (e.g., Relkin et al., 2020), most assessments report a single total score of overall CT ability. Although these assessments are valuable, they provide an *audit* of CT learning as opposed to granular evidence of children's CT understanding that can be directly linked to classroom instruction. In order to better support the development of CT in elementary classrooms, we propose using cognitive diagnostic models (CDMs), an approach to assessment that provides fine-grained information about what skills or attributes a child has or has not yet mastered. In CDMs, multiple latent concepts and skills, referred to as attributes, are identified and linked to assessment tasks indicating which attributes are required to solve each individual task. The scored responses of the tasks are used to provide a categorical classification (i.e., mastery vs. non-mastery) of the attributes. For example, if a child correctly responds to all the tasks that are linked to the attribute *identify bugs in buggy programs*, we infer they have mastery of the attribute. But what if they only answer half of those items correctly? Or what if a task is associated with more than one attribute? Knowing which attributes students have and have not mastered allows teachers to adjust classroom instruction based on their students' needs. The purpose of this paper is to explore how we can use CDMs by conducting an analysis on existing assessment data from a study where 271 children between the ages of 4 and 8 participated in a performance assessment, *Computational and Spatial Thinking* (CaST) assessment, designed to assess CT. Given the large range in ages, we are interested in whether performance is related to age. Our analysis was guided by the following questions: (1) *How does the CDM fit the performance assessment data? (2a) What mastery profiles of CT do children exhibit? (2b) Are these profiles associated with children's age?*

Assessment of computational thinking for emerging readers

As part of a larger project, we operationalized CT for early elementary classrooms and developed curricular resources and an assessment (CaST) around coding toys and coding environments that involve programming with directional codes *forward, backwards, rotate right, and rotate left* (see Clarke-Midura et al., 2021). Figure 1a shows children working on curricular tasks (on the left) and 1b shows a child taking the assessment and the materials associated with it (on the right). The development of the assessment was connected to and dependent on the development of the CT model and curricular tasks. We engaged in iterative cycles of design-based-research (DBR) where we refined each element (CT model, curriculum, assessment) based on new information learned in the process. We identified algorithmic thinking (AT), decomposition (modularity), debugging, abstraction, and spatial thinking (ST) as developmentally appropriate components of CT. We also identified mathematical knowledge (MK) that was required to solve CT tasks such as rotation on a point, linear units, and counting on.

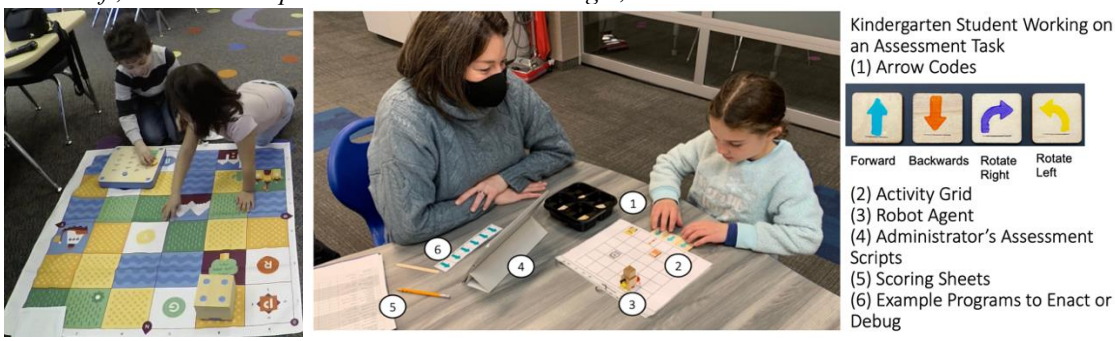
The Evidence Centered Design (ECD) framework (Mislevy & Haertel, 2006) guided the design of our CaST assessment tasks. The systematic process helped us articulate: the skills we wanted to assess, what inferences we wanted to support, what evidence we would need to support our inferences, the situations that would elicit the behaviors and observations of the skills and provide evidence, and how we would measure the skills. We used design patterns to document variable features of the tasks and the knowledge we thought each task was assessing. We specifically designed tasks to measure some of the skills we noted were necessary for CT tasks but were not part of most published CT models (e.g., rotation on a point and orientation of agent).

The CaST assessment is designed around a series of performance tasks (n = 19) that involve children either writing sequences of codes to navigate an agent from one location to another on a 6 × 6 2D grid, enacting programs by physically moving the agent on the grid, or debugging and fixing given programs using the four directional codes presented in Figure 1b (*forward, backward, rotate left, rotate right*). Given that the children we are assessing are emerging readers, the assessment is standardized and administered via a one-on-one format. The tasks are unplugged so the assessment can be used with a variety of coding toys and contexts that rely on navigational codes, which are common for pre-literate children. Some tasks have multiple correct answers, and all tasks were scored as incorrect or correct resulting in a total possible score of 19 points.

The assessment was validated in a prior study in which the items fit well to a two-parameter unidimensional Item Response Theory model (2PL IRT, see Na et al., 2023). The results of item analyses showed a high item discrimination (M = 2.26) and a moderate item difficulty (M = -.21), on average, with a high marginal reliability (r_{xx} = .87). IRT can estimate individuals' true ability score (θ) on a continuous scale, whereas CDMs classify examinees by whether or not they mastered each of the attributes that are required to successfully respond to the assessment tasks. Examinees are then classified into profiles based on the similarity of their responses. A benefit of CDM is that teachers can be provided with information on attribute mastery at both student and class levels.

Figures 1a and b

On the Left, Classroom Implementations and On the Right, Assessment Administration



Methods

Sample, procedures, and data source

Our sample included 271 children (girls = 142; aged 4-8; M_{age} = 6.54) from five elementary schools in the Western United States. For the analysis, age was categorized into three groups: young (< 72 months; n = 83), middle (72 ≤ months < 84; n = 104), and old (≥ 84 months; n = 84). The assessment was administered in a one-to-one interview format, by trained researchers in a quiet area in the schools. The administration took an average of 16.4 minutes per child. All assessments were video recorded and later scored by two independent researchers. Each task was scored as correct or incorrect. The two raters had high agreement (κ = .91). Tasks where there was no agreement were reviewed by the research team.

Data analysis

Statistical analysis was a multi-step process. The first step was to map the assessment tasks to a task-by-attribute table, which is called a Q-matrix. Identifying attributes entails hypothesizing what skills are needed to answer each task. We then validated and refined the Q-matrix. Our next step was to fit the Q-Matrix to the data by using a CDM model. We fit and compared three CDMs – DINA, DINO, and G-DINA – all of which have been widely adopted in empirical studies using CDMs. As a non-compensatory model, DINA (Deterministic Input, Noisy "AND" gate model) assumes that to answer a given task, children must possess all required attributes. For example, in the case of task 18 which is linked to two attributes, *fixing a program* (A2) and *spatial orientation of an agent*

(A3), the assumption under the DINA model is that a child must possess mastery of both attributes to successfully solve it. DINO (Deterministic Input, Noisy "OR" gate model) is a compensatory model, which assumes that if children have at least one attribute, they are likely to correctly respond to a task. In the case of task 18, if a child has mastered either *fixing a program* (A2) or *spatial orientation of an agent* (A3), they can correctly answer this task. Lastly, as a saturated model, G-DINA (Generalized Deterministic Inputs, Noisy "AND" gate model) assumes both compensatory and non-compensatory features within the test and therefore models the main effects of each attribute in conjunction with all possible interaction effects among attributes. Hence, in the example of task 18, children could have different levels of mastery probabilities depending on which attributes they have mastered or not. Both DINA and DINO are nested to G-DINA, which allows for log-likelihood test in model comparison. From the selected model, we conducted mastery profiles of each attribute, their classification accuracies, and identified mastery profiles to address our research questions. All statistical analyses were conducted in R (version 4.2.2) with *GDINA* package (Ma & de la Torre, 2020). Details of each step are described below.

Constructing, validating, and refining the Q-matrix

To construct a Q-matrix, we reviewed the assessment tasks and existing test specifications. The ECD process we used to design the assessment required that we document details of the tasks, such as variable features and knowledge being assessed, that we were able to use and share for identifying the attributes. We identified four attributes that were required to solve the tasks: *sequencing codes, fixing a program, spatial orientation of an agent, and rotation on a point* (see Table 1) and then mapped them onto the items into a Q-matrix.

Table 1

Four Attributes of the CT Assessment Tasks

	Attribute	Concept	Description
A1	Sequencing codes	AT	Represents the skill of ordering and arranging codes based on knowledge of syntax and semantics
A2	Fixing a program	Debugging	Represents the skill of implementing a successful strategy to fix bugs
A3	Spatial orientation of an agent	ST	Represents the skill of knowing that the codes always produce the same movements but depend on the agent's orientation
A4	Rotation on a point	MK	Represents the skill of knowing that a rotation occurs by rotating on a fixed point at a set angle, not translating to an adjacent point

Note. AT refers to algorithm thinking; ST refers to spatial thinking; MK refers to math knowledge.

Table 2

Refined Q-matrix for Four Attributes of CT and their PVAF values

Item	Attributes				PVAF	Item	Attributes				PVAF
	A1	A2	A3	A4			A1	A2	A3	A4	
1	1	0	0	1	.941	10	0	1	0	1	.988
2	1	0	0	0	.949	11	0	1	0	1	.990
3	1	0	1	0	.994	12	0	1	1	1	.999
4	0	1	0	0	.862	13	0	1	1	1	.991
5	0	0	0	1	.966	14	1	0	1	0	.933
6	0	1	0	1	.999	15	1	0	1	0	.898
7	1	0	1	1	.984	16	1	0	1	0	.981
8	0	1	1	1	.994	17	1	0	1	0	.974
9	0	1	0	1	.997	18	0	1	1	0	.993

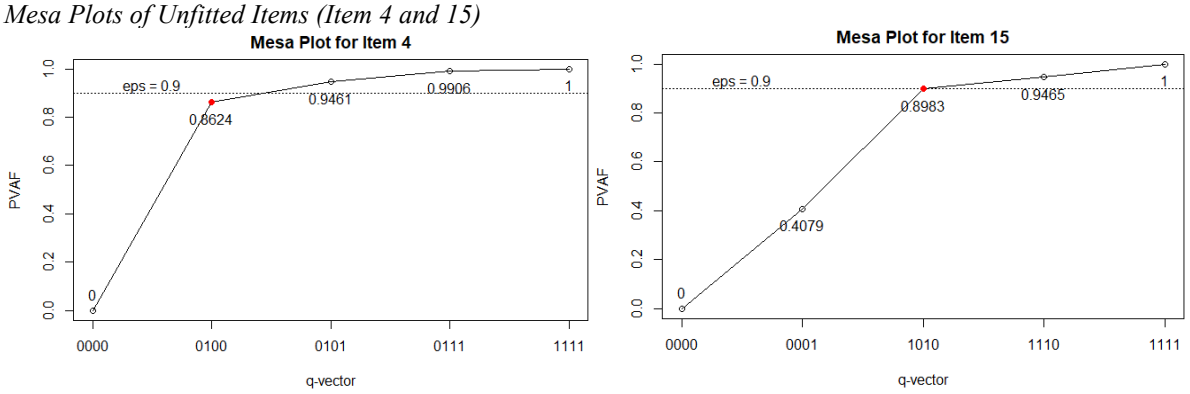
Note. "1" refers to required attributes to solve given items, whereas "0" indicates non-required in the attribute of each item. A1 refers to sequencing of codes, A2 refers to fixing a program, A3 refers to spatial orientation of an agent, and A4 refers to rotation on a point. *PVAF* refers to the proportion of variance accounted by q-vectors.

The Q-matrix was qualitatively validated by expert review of two raters ($\kappa = .82$). In order to validate the Q-matrix quantitatively, the proportion of variance of accounted (PVAF, de la Torre & Chiu, 2016) by each item-attribute specification (i.e., q-vector) was calculated and the acceptable PVAF values set to .90. We also used a mesa plot (see Figure 2), which visualized the relationship between possible q-vectors in the x-axis, and PVAF values in the y-axis to visually investigate possibilities to refine the Q-matrix. Fitting the initial Q-matrix to the assessment response data (19 tasks) did not yield acceptable model fits, so we eliminated one task and modified q-vectors of five tasks. After iterative modifications of the Q-matrix, we refitted the refined Q-matrix to

the response data (18 tasks), resulting in acceptable model fits. While most tasks showed acceptable PVAF values, a mesa plot suggests that task 4 (PVAF = .862) and task 15 (PVAF = .898) need further modifications of q-vectors (see Figure 2). However, we did not modify q-vectors of these two tasks because these changes led to only minuscule increases in PVAF (less than .10 of changed PVAF) and were not aligned with what these tasks intended to measure from expert reviews. As a result, we used data from 18 of the 19 tasks with the refined Q-matrix (see Table 2) in which three tasks were assigned to one CT attribute, eleven tasks were assigned to two CT attributes, and four tasks were assigned to CT three attributes.

Figure 2

Mesa Plots of Unfitted Items (Item 4 and 15)



Note. X-axis refers to item-attribute specifications (q-vectors) and Y-axis refers to PVAF refers to proportion of the variance accounted for the q-vectors. A filled dot in the mesa plot means the q-vector denoted in the Q-matrix, and black dots mean possible q-vectors in the Q-matrix. Eps (epsilon) refers to a designated threshold value of PVAF.

Selecting the appropriate CDMs

In order to select the most appropriate model, we evaluated the model fits of the three models (see Table 3). We specifically looked at the Akaike Information Criterion (AIC) and the likelihood ratio test (LRT). As shown in Table 3, G-DINA showed the lowest AIC and the LRT was significant when comparing the general model (G-DINA) to the reduced models DINA (LR: 205.11, $df = 46$, $p < .001$) and DINO (LR: 232.80, $df = 46$, $p < .001$). Thus, G-DINA was selected for the CDM model for subsequent analyses.

Table 3

Model Fit Indices for G-DINA, DINA and DINO

Model	AIC	*n*Pars	Loglik	Likelihood Ratio Test		
				LR	*df*	*p*-value
G-DINA	4635.68	97	-2220.84			
DINA	4748.79	51	-2323.40[a]	205.11	46	<.001
DINO	4674.47	51	-2337.24[b]	232.80	46	<.001

Note. AIC refers to Akaike information criterion; *n*Pars refers to number of model parameters; Loglik refers to log likelihood; LR refers to likelihood ratio; [a]G-DINA versus DINA; [b]G-DINA versus DINO.

To address RQ 1, using the G-DINA model, we estimated mastery probabilities and classification accuracies for each attribute. Using expected a posteriori (EAP, Huebner & Wang, 2011), children were classified as mastery of attributes ("1") when their mastery probabilities of each CT attribute were above .50; otherwise, they were classified as non-mastery ("0"). For example, if a child has .373 for *sequencing of codes* (A1), .829 for *fixing a program* (A2), .992 for *spatial orientation of an agent* (A3), and .171 for *rotation on a point* (A4), their mastery status of each attribute is "0" for *sequencing of codes* (A1), "1" for *fixing a program* (A2), "1" for *spatial orientation of an agent* (A3), and "0" for *rotation on a point* (A4), resulting in a mastery profile of "0110". Mastery proportions of each attribute – the ratio of the number of children who have mastered a given attribute to the total number of the sample - represent their relative difficulty, and their classification accuracies indicate the reliability of classifying children's mastery status as either mastery or non-mastery.

To address RQ 2, we estimated individuals' mastery profiles of the four CT attributes from the CDM results. For example, a mastery profile of 0100 refers to a set of children who have mastered *Fixing a program* (A2) but have not yet mastered the other three CT attributes. We evaluated which mastery profiles were common

or rare among our sample. We further conducted a chi-squared test of independence to examine the associations between identified mastery profiles and age groups. Results of this analysis are presented below.

Results

The viability of classifying children's mastery status of CT attributes

As we see in Table 4, approximately 48% of children mastered *sequencing of codes* (A1) and *fix a program* (A2), respectively. Fewer children mastered *spatial orientation of agent* (A3; 37.3%) and *rotation on a point* (A4; 40.6%). While mastery proportions of four attributes were higher for the older children (see Figure 3), we note that mastery proportions of *spatial orientation of an agent* (A3) were substantially lower (59.5%) than the other three attributes in older children. Likewise, in middle age children, *spatial orientation of an agent* (A3, 27.9%) and *rotation on a point* (A4, 26.0%) showed lower mastery proportions than the other two attributes. The estimated classification accuracies at the attribute level were high; they ranged from .90 to 97 and were .85 at the test-level. These values the G-DINA model reliably classifies children into attribute mastery. Figure 3 presents the mastery proportions of each attribute within the total sample and by each age group.

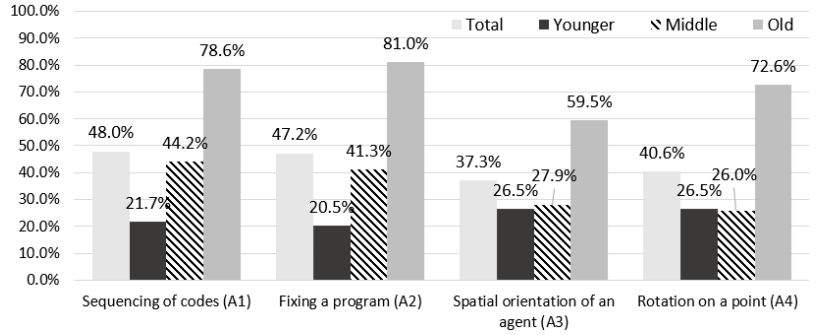
Table 4

The Proportion of Mastery of Four CT Attributes and their Classification Accuracies

	A1. Sequencing of codes	A2. Fix a program	A3. Spatial orientation of an agent	A4. Rotation on a point
Mastery proportion (%)	48.0%	47.2%	37.3%	40.6%
Classification accuracy	.96	.96	.90	.97

Figure 3

Mastery Proportions of Each Attribute in the Total Sample and Age Groups

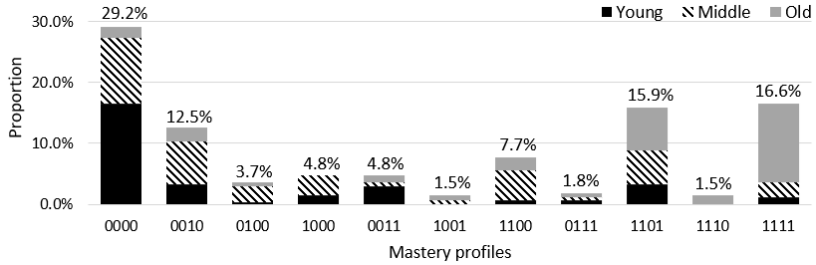


Detecting CT mastery profiles and the profiles association with children's age

CDM results yielded 11 mastery profiles out of a possible 16. Figure 4 shows the distribution of the profiles by age group where 0000 means that none of the four attributes were mastered and 1111 means that all of the four attributes were mastered. A chi-squared test of independence confirmed that the mastery profiles are statistically associated with the age groups, $\chi^2(20) = 123.83$, $n = 271$, $p < .001$, Cramer's V = .478.

Figure 4

Identified 11 Mastery Profiles of CT Components by Age Groups



We focus on the four most common profiles (those with proportions > 10%):

- *Non-mastery profile* (0000, $n = 79$). This profile represents children who have not mastered any of the CT attributes. It is the most common profile (29.2% of children). It is comprised of mostly young (n =

45) and middle (n = 29) children, compared to a small number of the older children (n = 5). This profile had the lowest total scores on CaST assessment among all 11 mastery profiles, M = 3.09, SD = 1.53.

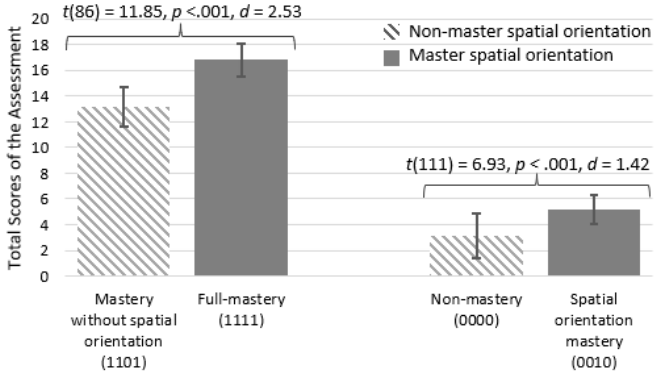
- *Full-mastery profile* (1111, n = 45). This represents children who have mastered all four CT attributes. It is the second most common profile (16.6%). The group is comprised primarily of older children (n = 35) and some middle (n = 7) and younger (n = 3) children. This profile scored the highest CaST total score, M = 16.80, SD = 1.10.
- *Mastery without spatial orientation profile* (1101, n = 43). This profile represents children who mastered all of the CT attributes except *spatial orientation of an agent* (A3). It includes 15.8% of the sample and was comprised mostly of older children (n = 19) and middle children (n = 15) compared to younger children (n = 9). This profile had relatively high total scores on CaST assessment, M = 13.14, SD = 1.74.
- *Spatial orientation mastery profile* (0010, n = 34). This group only mastered *spatial orientation of an agent* (A3). It includes 12.5% of the sample. This profile was comprised of middle (n = 19), young (n = 9) and older (n = 6) children. This profile had relatively low total scores on CaST assessment, M = 5.18, SD = 1.29.

Importance of spatial orientation of an agent

Based on the attribute mastery results, we decided to conduct an ancillary analysis to explore the role of *mastering spatial orientation of an agent* (A3) on overall CT abilities. Using the total assessment score as a proxy of overall CT abilities, we conducted independent t-tests between the *full-mastery* profile (1111) and *mastery without spatial orientation* profile (1101); and between the *non-mastery* profile (0000) and s*patial orientation mastery* profile (0010) to check the role of spatial orientation in the overall CT ability. The result of the independent t-tests showed that the *full-mastery* profile (M = 16.80; SD = 1.10) significantly outperformed the *mastery without spatial orientation* profile (M = 13.14, SD = 1.74) on the CaST assessment, based on the CaST total score, $t(86) = 11.85$, $p < .001$, d = 2.53 (see Figure 5). The *spatial orientation mastery* profile (M = 5.18, SD = 1.29) had significantly higher CaST total scores than the *non-mastery* profile (M = 3.09, SD = 1.53), $t(111) = 6.93$, $p < .001$, d = 1.42 (see Figure 5). The results suggest that mastering *spatial orientation on an agent* (A3) played a significant role in children's overall CT abilities.

Figure 5

Mean Differences in the CaST Total Scores by whether Children Mastered Spatial Orientation



Discussion

This study explored the viability of using CDMs to diagnose mastery levels of children's CT abilities on a finer grain size by looking at four attributes: *sequencing of codes, fixing a program, spatial orientation of the agent, and rotation on a point*. We were able to fit a CDM model, G-DINA, to the CaST assessment response data and yielded information about the mastery proportions of the four attributes as well as profiles of attribute mastery in the sample, by participant's age. We conducted an additional analysis based on our results to explore the role of spatial orientation on overall CT ability using the CaST assessment total score as a proxy.

We hypothesized that knowledge of sequencing of codes, fixing a program, spatial orientation of the agent, and rotation on a point were required for children to answer the tasks on the CT assessment. Looking at just the mastery of attributes, our results indicate that higher proportions of older children mastered the four CT attributes and that the proportion of older children who mastered spatial orientation of an agent was lower than for the other three CT attributes (Figure 3). Overall, much smaller proportions of middle and younger children

mastered all four attributes. Smaller proportions of middle children mastered spatial orientation of an agent and rotation on point. Yet for the younger children greater proportions showed mastery of spatial orientation of the agent and rotation on a point than the sequencing of codes and fixing a program attribute. Overall, the older children had higher proportions of mastery in the CT-related concepts than the spatial and mathematical concepts whereas the younger children had higher proportions of mastery in the spatial and mathematical related concepts. These findings align with the results of RQ 2a that show the most common mastery profile was the *Non-mastery profile*. The mastery profiles show how children's responses and attribute mastery cluster into patterns. Focusing on the four most common mastery profiles, we see that the old and middle age children were likely to be assigned to the *full-mastery* and *mastery without spatial orientation profile*, whereas the young age group were more likely to be assigned to the *non-mastery* profile. Put differently, the young children in our sample (< 72 months) have not yet mastered most of the attributes of CT, while some of the children in the middle and older age groups (≥ 72 months) either mastered all the CT attributes or needed only more experiences with spatial orientation of the agent. These findings align with Relkin et al. (2020) who in a sample of a similar age range of children found that older children performed better on measures of CT. These findings suggest that children's understanding and proficiency in CT may be associated with their age, which supports the need not only for developmentally appropriate curriculum, resources, and assessments for fostering and measuring CT in early childhood but a need to provide younger children with opportunities to engage with CT through playing with coding toys.

Perhaps the most interesting mastery profile is the *spatial orientation mastery profile* (0010, n = 34). This group only mastered one attribute: *spatial orientation of an agent* (A3). While it only included 12.5% of the sample, it was mostly comprised of middle age children (n = 19), with some young (n = 9) and older (n = 6) children. This profile had relatively low total scores on the CaST assessment (M = 5.18, SD = 1.29). We conducted additional analyses to explore the role of the spatial orientation of an agent attribute on CT knowledge, using the total CT assessment score as a proxy. The results suggest that mastering *spatial orientation on an agent* (A3) played a significant role in the overall CT abilities, as measured by the overall score on the CT assessment. Spatial thinking (ST) entails understandings of space and objects' positions in space, reasoning with objects or representations in space, and operations on spatial relationships (NRC, 2006). A component of ST, spatial orientation is the understanding of different positions in space, and children first develop spatial orientation concepts in relation to their own position in space and later develop external-based reference systems using landmarks outside themselves (Sarama & Clements, 2009). Researchers have identified a number of factors constituting spatial thinking skills; however, there is no consensus in its exact structure or consistency in measurement (Atit et al., 2020). Existing research on the relationship between CT and ST has looked at children's relation of CT skills with non-verbal visuospatial reasoning (Tsavara et al., 2022), mental rotation skills (Città et al., 2019), and spatial ability (Román-González et al., 2017) and found significant correlations between concepts of ST and CT in early childhood. Our findings further support the importance of the relationship of ST and CT and the need to better understand this relationship in early childhood.

The older and middle age group of children in our sample had higher mastery probabilities of sequencing codes and debugging programs, than spatial orientation of the agent (see Figure 3). It could be that the kinds of exposure to CT that the older children have through coding provide more experience with practices like sequencing and debugging and less with spatial orientation of agents. In the US, kindergarten standards mostly focus on applying spatial knowledge as relational from their own perspective and not from different perspectives, which means children do not get a lot of exposure to this in kindergarten. Previous research on young children playing with tangible coding toys observed children shifting back and forth between egocentric and allocentric perspectives, or reference frames, while programming robots to navigate paths on the floor. Children's inability to take on an allocentric perspective, the robot's perspective when the robot was facing a different orientation often resulted in coding errors such as selecting the wrong code (Clarke-Midura et al., 2021; Wang et al., 2021). Our findings support these findings and suggest the importance of playing with tangible coding toys at a young age to aid in the development of both ST and CT skills. Finally, research has shown that ST is a critical component of STEM learning and practices and is domain dependent (Atit et al., 2020). While ST skills are malleable and can be improved through training and instruction (Uttal et al., 2013), instead of fostering ST independently, it is more critical to situate ST into overall CT learning activities. As mentioned above, more research is needed that explores the relationship of ST and CT skills in early childhood.

Limitations and conclusion

Despite the multi-faceted nature of CT, we only selected four CT attributes due to our limited sample size. There is a need for future studies that include larger samples and more attributes, especially those related to ST and MK. Nevertheless, identifying which CT attributes children have not mastered as well as what attributes are foundational to CT learning is an important step toward designing and implementing tailored learning experiences,

and minimizing potential gaps in CT and STEM learning from an early age. Furthermore, there is a need for developmentally appropriate curricular resources and assessments of CT for early childhood. CDMs offer a potential way to provide teachers with informative information about their students' CT understanding that they can directly link to their instruction. As a field, learning scientists tend to talk about the design of assessments and learning environments separately. The present study shows the affordance of thinking about instruction, assessment, and theories of learning as an integrated system.

References

Atit, K., Uttal, D. H., & Stieff, M. (2020). Situating space: Using a discipline-focused lens to examine spatial thinking skills. *Cognitive Research: Principles and Implications, 5*(1), 1-16.

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. *Computers & Education, 141*, 103613.

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.

de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253–273.

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement, 71*(2), 407–419.

Luo, F., Israel, M., & Gane, B. (2022). Elementary computational thinking instruction and assessment: A learning trajectory perspective. *ACM Transactions on Computing Education (TOCE), 22*(2), 1-26.

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software, 93*(14), 1–26.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6-20.

Na, C., Clarke-Midura, J., Shumway, J. F., Silvls, D., & Lee, V. R. (2023). *Validating a performance assessment of computational thinking for early childhood using Item response theory*. Manuscript submitted for publication.

National Research Council. (2006). *Learning to think spatially*. The National Academies Press.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

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*(4), 482-498.

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.

Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children*. Routledge.

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. *Computers & Education, 179,* 104425.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: a meta-analysis of training studies. *Psychological Bulletin, 139*(2), 352-402.

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 de Vries, E., Hod, Y., & Ahn, J. (Eds.), *Proceedings of the 15th International Conference of the Learning Sciences - ICLS 2021.* (pp. 394-401). Bochum, Germany: International Society of the Learning Sciences.

Wing, J. (2014). *Computational thinking benefits society*. Retrieved from http://socialissues.cs.toronto.edu/

Acknowledgement

This work was supported in part by funding from the National Science Foundation under Grant No. DRL-1842116. The opinions expressed herein are those of the authors and do not necessarily reflect those of the National Science Foundation.