Design and validation of learning trajectory-based assessments for computational thinking in upper elementary grades

Brian D. Gane, Ph.D.
Learning Sciences Research Institute, University of Illinois at Chicago, Chicago, IL, USA. ORCiD: https://orcid.org/0000-0002-4454-569X

Maya Israel, Ph.D.
Educational Technology, University of Florida, Gainesville, FL, USA. Twitter: @misrael09 ORCiD: https://orcid.org/0000-0003-0302-6559

Noor Elagha
Psychology, University of Illinois at Chicago, Chicago, IL, USA.

Wei Yan
Educational Technology, University of Florida, Gainesville, FL, USA. Twitter: @WeiYan2018

Feiya Luo, Ph.D.[1]
Educational Technology, University of Florida, Gainesville, FL, USA. Twitter: @feiyaluo ORCiD: https://orcid.org/0000-0002-3037-085X

James W. Pellegrino, Ph.D.
Learning Sciences Research Institute, University of Illinois at Chicago, Chicago, IL, USA. linkedin.com/in/james-pellegrino-02b21a4

Corresponding author: Brian Gane. Email: bgane@uic.edu

Biographical notes

Brian D. Gane is a research assistant professor at the Learning Sciences Research Institute at the University of Illinois at Chicago. His primary research interests center around the research and development of learning environments, including the design of assessments, instruction, and curriculum within those learning environments. In particular, he focuses on learning in science and engineering disciplines, especially using assessments to measure and support student learning. Additionally, he collaborates with K-12 teachers to study methods to develop assessment literacy and practices, including how to use assessment formatively to advance students' learning.

[1] Dr. Feiya Luo is now an assistant professor of Instructional Technology in the College of Education at the University of Alabama.

Maya Israel, Ph.D. is an associate professor in the Department of Educational Technology in the School of Teaching and Learning at the University of Florida. She is also the research director at the Creative Technology Research Lab. Her research focuses on strategies for supporting learners' meaningful engagement in science, technology, engineering, and mathematics (STEM) with emphases on computer science education and Universal Design for Learning (UDL). She is currently PI on an NSF STEM+C grant to study learning trajectories that align computational thinking with elementary math instruction. She is also PI on an NSF grant that examines instructional strategies that support struggling learners in successfully accessing computing instruction. Dr. Israel was a writer on the Framework for K-12 Computer Science Education and the revisions of the Computer Science Teachers Association (CSTA) Standards for Computer Science Teachers. She has published in journals such as *Computer Science Education, Exceptional Children, Journal of Research on Technology in Education, Journal of Research in Science Teaching,* and *Computers & Education* as well as in computer science education conferences such as SIGCSE and ICER. Lastly, Dr. Israel also works with multiple school districts on meaningfully including all learners in computer science education initiatives.

Noor Elagha is a graduate student in Cognitive Psychology at the University of Illinois at Chicago and research assistant at the Learning Sciences Research Institute at the University of Illinois at Chicago. Her research interests include learning processes and mathematics learning interventions.

Wei Yan is a Ph.D. student in Educational Technology at the University of Florida. Her research interests are computer science education, computational thinking integration in K-8 education, and computer-supported collaborative learning.

Feiya Luo graduated with a Ph.D. in Educational Technology from the University of Florida. She is an assistant professor in the College of Education at the University of Alabama. Her research interests include elementary computational thinking integration and computer science education.

James W. Pellegrino is Liberal Arts and Sciences Distinguished Professor and Co-director of the Learning Sciences Research Institute at the University of Illinois at Chicago. His research and development interests focus on children's and adult's thinking and learning and the implications of cognitive research and theory for assessment and instructional practice. He has published over 300 books, chapters and articles in the areas of cognition, instruction and assessment. He has served on several National Academy of Sciences study committees, including chair of the Committee for the *Evaluation of the National and State Assessments of Educational Progress*, co-chair of the Committee on the *Foundations of Assessment,* chair of the Committee on *Defining Deeper Learning and 21st Century Skills,* and co-chair of the Committee on *Developing Assessments of Science Proficiency in K-12*. He is a past member of the Board on Testing and Assessment of the U.S. National Research Council, and a lifetime member of the National Academy of Education and the American Academy of Arts and Sciences.

Design and validation of learning trajectory-based assessments for computational thinking in upper elementary grades

Background & Context: We describe the rationale, design, and initial validation of computational thinking (CT) assessments to pair with curricular lessons that integrate fractions and CT.

Objective: We used cognitive models of CT (learning trajectories; LTs) to design assessments and obtained evidence to support a validity argument.

Method: We used the LTs and evidence-centered design to develop assessments that 144 Grade 3 and Grade 4 elementary students completed following the integrated instruction. We analyzed data using multiple psychometric approaches.

Findings: The design approach and data analysis suggest that the assessments are well-suited to evaluate students' CT knowledge, skills and abilities across multiple LTs.

Implications: We show how to use LTs to design assessments that can yield valid inferences about students' CT competencies; these methods can be adopted and extended by others to create additional assessments that can advance computer science education.

Keywords: learning trajectories; assessment; computational thinking; evidence-centered design; validity; elementary

Introduction and background

Computer science (CS) education has rapidly expanded into the elementary grades. The rationale for this expansion often includes assertions that exposing elementary students to CS and computational thinking (CT) helps to develop their problem-solving skills (e.g., Voogt et al., 2015; Voskoglou & Buckley, 2012) and persistence during challenging tasks (e.g., Barr et al., 2011; Barr & Stephenson, 2011). This literature further suggests that engaging in these early and sustained computational experiences provides students an advantage given the ubiquity of computational technologies. Significant efforts are currently underway to develop a wide range of methods of assessing young learners' CT knowledge and skills in response to this expansion. These efforts include developing and using assessments that utilize automated tools such as Dr. Scratch (e.g., Moreno-León et al., 2015), Parson's puzzles (e.g., Denny et al., 2008; Ericson et al., 2017), rubrics for evaluating computational artifacts (e.g., Seiter & Foreman, 2013), artifact-based interviews (e.g., Brennan & Resnick, 2012), and paper-and-pencil assessments of CT concepts (e.g., Basawapatna, 2011; Román-Gonzalez et al., 2017). While many CT assessments have been designed, only a small number of studies provided validity evidence (Tang et al., 2020), making it difficult for researchers and teachers to interpret existing CT assessments with confidence. In addition, most assessments lack a theoretical framework that consistently identifies which CT constructs are assessed and at what levels of sophistication. Therefore, it is crucial to use an assessment design approach that is theoretically grounded and guided by well-established design frameworks in order to fully unravel the complexities of students' CT skill development and to capture students' CT understanding (Tang et al., 2020). In this paper, we describe a CT assessment development effort that is theoretically grounded in three literatures: (1) CT literature, through the use of CT learning trajectories to specify the model of student knowledge, (2) assessment design literature, through the use of an evidence-centered design logic, and (3) assessment validation literature, using an argument-based approach for specification, analysis, and interpretation of relevant data.

Given that elementary CT education is still new, efforts to develop and implement CT assessments are taking place at the same time that researchers and practitioners are debating the definitions of CT, experimenting with creating developmentally appropriate instructional materials, and exploring ways to integrate CT into other content areas (e.g., mathematics and science). These complexities must be acknowledged in any discussion of developing, validating, and implementing CT assessments as they have significant implications for their design and use. In fact, development of CT assessments can advance these efforts because measurement of hypothetical constructs like CT provides an operational definition of the construct (Croker, 2006), as necessary for researchers as it is for practitioners.

Definitions of CT

CT is often defined as a mental process for formulating and solving problems in a manner that draws on concepts that are fundamental to computer science (Wing, 2006). This definition has been expanded and revised numerous times (Barr et al., 2011), with no single definition emerging to guide the development of CT instruction and assessment. In all likelihood, there is no single definition of CT that applies across all content areas and grade levels. However, Shute and colleagues (2017) synthesized CT definitions across multiple studies and frameworks and came up with a consolidated definition of CT, "The conceptual foundation required to solve problems effectively and efficiently (i.e., algorithmically, with or without the assistance of computers) with solutions that are reusable in different contexts" (p. 151). Although outside the scope of this paper, interested readers can refer to Shute et al. (2017) for an in-depth analysis of the competing definitions of CT.

The K-12 CS Framework (2016) and the revised CSTA standards (2017) define five computational concepts (i.e., computing systems, networks and the internet, data and analysis, algorithms and programming, and impacts of computing) and seven practices (fostering an inclusive computing culture, collaborating around computing, recognizing and defining computational problems, developing and using abstractions, creating computational artifacts, testing and refining computational artifacts, and communicating about computing). These concepts and practices showcase the wide scope of CS and the interrelatedness of CT. CS instruction should not be considered synonymous with programming. Moreover, although CT is also more than programming, programming can be an important tool for developing CT (Voogt et al., 2015). Elementary CS and CT instruction often emphasizes algorithms and programming with activities such as developing precise algorithms that computers can interpret, decomposing complex tasks into simpler tasks, and sequencing instructions either sequentially or by repeating patterns of instructions or by using events to initiate instructions (K-12 CS Framework, 2016). Although not the only computational concepts taught in elementary CS, algorithms and programming are often addressed as a means of practicing problem-solving.

CT integration

Another element that adds complexity for assessing CT in the elementary grades is that CT learning often occurs within an integrated context (e.g., along with science or mathematics learning). The rationale for integration includes (a) the natural "fit" of CT into certain disciplines (Lee et al., 2014; Weintrop et al., 2016), (b) opportunities to deepen students' understanding of content knowledge (Bers et al., 2014; Yadav et al., 2016), (c) limited time in elementary schedules for new content areas (English, 2017), and (d) promoting equity by providing CT within content areas taught to all learners (Israel et al., 2015; Santo et al., 2019). As part of a larger project, we have been developing curricular lessons that integrate CT within mathematics for students in Grades 3 and 4 (Strickland et al., 2021) and assessments to complement these lessons.

Given integrated approaches to CT instruction at the elementary level, CT assessments must be developed to measure CT learning goals while being cognizant of the specific instructional context in which CT is introduced. This challenge is not unique to CT education, as it often occurs when assessing learning in integrated science, technology, engineering, and mathematics (STEM) education (e.g., Douglas et al., 2020; Harwell et al., 2015). In this paper, we describe the approach we adopted that emphasizes assessing students' CT knowledge, skills, and abilities within the context of curricular lessons that integrate CT into mathematics. As such, we are interested in making claims about students' CT competencies; we are not trying to simultaneously measure both their CT and math knowledge, skills, and abilities by using the same assessments. Instead, we focus on developing assessments from the same CT foundation as used to develop lessons. Many of our CT assessments do involve mathematics, but do so to provide a context or a problem to be solved, rather than as a means to measure students' math competencies.

Learning trajectories to inform material development

At this time, there is no systematic, evidence-based means to decide *how* elementary students should learn CT (Zhang & Nouri, 2019). In the absence of such empirical evidence, our decision making relied on six learning trajectories (LTs) that were developed based on implicit goals in research studies in the areas of sequence, repetition, conditionals, decomposition, variables, and debugging (Rich et al., 2017, 2018, 2019, 2020).

LTs have three primary components: goals, developmental progressions of students' thinking processes, and instructional activities that promote the development of students' thinking along the progressions (Clements et al., 2020; Maloney et al., 2014). When LTs are used to guide lesson development, lessons are designed to build upon contiguous levels of instruction in a deliberate, consecutive manner because those lessons build on foundations set by previous lessons in order to avoid gaps in learning. If instruction is based solely on CT target competencies (e.g., lessons on sequencing), we assume that students will be able to perform tasks associated with that level of instruction as well as all previous levels.

The CT LTs used in this study were previously developed after analyzing CS/CT education peer-reviewed articles that were categorized by the learning goals identified in those studies. These learning goals were then organized into consensus goals (CGs) that included both an understanding goal (i.e., what students should know) and an action goal (i.e., what students should be able to do). These LTs were designed with the purpose of beginning to understand how students learn CT. As such, it is important to acknowledge that although these LTs may appear to be organized in a linear manner, learning does not occur in such a linear manner. Additionally, these LTs are considered hypothetical as empirical evidence for these LTs is only beginning to emerge as researchers begin to test students' CT development (e.g., Luo, 2020); doing so requires assessments that can be aligned to the LTs. Lastly, each LT does not exist in isolation; rather, they are connected. As students progress through one LT, progress on other LTs may be required or supported. For example, the sequence goal, *Computers require precise instructions using limited commands* can naturally lead into the repetition goal, *Instructions like "Step 3 times" do the same thing as "step, step, step"* and the decomposition goal, *Systems are made up of smaller parts.* These goals can lead into a conditional goal such as, *Actions often result from specific causes*. Thus, each LT is distinct, but increasing knowledge in one LT supports understanding in other LTs. More information about the development and features of the CT LTs are available in several papers by Rich and colleagues (2017, 2018, 2019, 2020) and on our website ([BLIND]).

Figure 1 depicts a portion of the repetition LT. This LT provides a series of learning goals for how elementary students can begin with a common understanding of repetition (i.e., *2:*

Some tasks involve repeated actions) as a starting point for understanding iteration in everyday life (i.e., *1: Repeating things can have a cumulative effect*), and the utility of loops in programming (i.e., *5: computers use repeat commands*). Each learning goal is connected to others to indicate progressions in knowledge, skills, and abilities. Our lessons were designed to guide students to build on each learning goal's level of thinking.

Curricular lesson development

Although the CT LTs extend beyond 3rd and 4th grade, our team focused on CT learning in 3rd and 4th grade fractions instruction, which were designed to guide learners in achieving the knowledge, skills, and abilities described in the six CT LTs. Thus, a series of 11 3rd grade and 12 4th grade lessons were developed to replace or supplement instruction available through the *Everyday Mathematics* (4th ed.; EM4) curriculum (Strickland et al., 2021). EM4 was chosen as it aligns to the Common Core State Standards for Mathematics (CCSS-M) and is a commonly used elementary curriculum in local school districts. Within the EM4 curriculum, we focused primarily on fractions instruction for two reasons. First, many students find proportional reasoning involving fractions difficult (Boyer & Levine, 2012). Second, CT and fractions instruction appear to be complementary. For example, many fraction standards in Grades 3 and 4 are framed around the notion that unit fractions can be combined repeatedly to produce non-unit fractions, which is related to the idea of repetition and looping in CT.

The 3rd and 4th grade lessons included a combination of unplugged activities (i.e., lessons that do not require interactions with the computer) and plugged activities (i.e., lessons that included computer-based activities). For plugged activities, we utilized the Scratch programming environment—a visual block-based programming environment designed for young learners (Maloney et al., 2010)—for several reasons. First, it is one of the best known programming environments for teaching CT to young learners (Price & Barnes, 2015). In fact, as of October, 2020, Scratch is the highest block-based programming language on the TIOBE index (https://www.tiobe.com/tiobe-index/), which measures search engine results of the popularity of programming languages. Second, Scratch is supported on Chromebooks, which were the computers available in our research schools. Most importantly, Scratch was chosen for pedagogical and theoretical reasons. Scratch was designed using a constructionist viewpoint that encourages young learners with no previous programming experience to explore, learn, and share (Maloney et al., 2010). It was designed based on Papert's assertion that programming environments should have a "low floor" (i.e., easy to start) and a "high ceiling" (i.e., options to increase complexity; Resnick et al., 2009).

Regardless of whether the lessons were plugged or unplugged, lesson design used the same guided instruction approach aligned to the EM4 lesson design. All lessons started with a "warm up" that activated prior knowledge, made instructional learning goals explicit, and provided an introduction to the new content. This was followed by the "focus" portion, the main instructional time, and ended with a wrap up to revisit goals and reflect on learning (Strickland et al., 2021).

Assessment development and field testing

As with the curricular lessons, we used the LTs to develop assessment items to measure students' CT knowledge and skills. There are three main features of the assessment items. First, all assessments used a "paper-and-pencil" format, so they did not require interactions on a computing device. Although they were paper-and-pencil, some assessment items used the Scratch interface. These items included illustrations of Scratch code blocks (and/or the Scratch interface) and/or asked students to draw Scratch code blocks by hand (e.g., Figure 2). Other assessment items were designed to elicit students' CT without being embedded within the Scratch programming environment (e.g., Figure 3). Second, some items used a mathematics context while others did not. However, the mathematics content was simplified for the assessment items to reduce the likelihood that students' proficiency with the focal mathematics content would preclude them from answering the items in a way that would demonstrate their CT proficiency. For instance, the item in Figure 2 asks students to move the cat along the number line which uses integers. In the lesson, students complete a similar task in which the number line uses fractions. By using integers in the assessment task, students without a full understanding of fractions could still complete this task (in the *Everyday Mathematics* curriculum, students have experience using a number line with integers in a prior grade). Third, all assessment items were designed as measures of a single LT topic (e.g., repetition). Thus, the produced assessment items could be classified along three dimensions:

1. Whether the item does (or does not) use the Scratch interface,
2. Whether the item is (or is not) embedded in a mathematics context, and
3. Which LT (topic) is the primary focus for that item.

Román-González et al. (2019) classify CT assessment tools based on their method of evaluation: diagnostic tools, summative tools, formative–iterative tools, data-mining tools, skill transfer tools, perceptions–attitudes scales, and vocabulary assessment. Our assessments are designed as tools to measure near transfer that can be used diagnostically and/or summatively.

Theoretical and design frameworks

In this section, we describe the theoretical and design frameworks used to develop the assessments and interpret the assessment results. Underlying assessment use is the principle that the assessment must be valid with respect to the intended interpretive use. The design and interpretation of assessments are intimately intertwined and jointly affect valid use of assessments. Therefore, before describing the development approach and interpreting results, we frame our assessments within theoretical understandings of validity. We use the assessment triangle (Pellegrino, 2001) to structure the chain of reasoning between interpretation and use. To systematically apply the concepts represented by the assessment triangle, we use evidence-centered design. Each of these theories and design frameworks are described in turn.

Validity

In their systematic review of CT assessments, Tang and colleagues (2020) conclude that across almost 100 studies, only 18% reported on the validity of their assessments. Those that did consider validity tended to focus on criterion, content, and construct validity. In this paper, we follow a different tack, focusing on contemporary approaches that conceptualize validity as an argument (AERA, APA, & NCME, 2014; Douglas et al., 2015). In an argument-based approach to validity (Kane, 1992; 2006), one makes claims about the intended interpretive use of the assessment and then provides evidence in support of those claims. This evidence can and should be multifaceted, including information about the design of the assessments as well as empirical evidence from student performance (Crocker, 2006; Pellegrino et al., 2016). After instruction with the curricular lessons, students should be able to demonstrate CT as they solve items that were designed to recruit specific aspects of CT. Specifically, they should be able to use the knowledge, skills, and abilities that are developed during instruction and enactment of the lessons.

For assessments that are intended to support instruction, there are three components of validity that can be considered: instructional, inferential, and cognitive; each have their own associated claims and potential sources of evidence (Pellegrino et al., 2016). This paper focuses on providing preliminary evidence for claims related to the cognitive and inferential components of validity. The cognitive component is centered on the assessment items being an accurate measure of students' CT as articulated by the LTs, and in this paper is evidenced by the assessment design methodology. The inferential component is related to whether the assessments reliably differentiate among students and support model-based analyses of their measurement properties and in this paper is evidenced by results from the analysis of students' performance on the assessments.

We have one primary claim that we advance and support in this paper, namely that **the CT assessment items measure students' performance with respect to the LTs (topics) that students were introduced to during the lessons**. This primary claim has three constituent subclaims:

A. Items are aligned to the LTs and students' opportunities to learn,
B. Item scores are reliable (i.e., item scoring rules and rubrics can be used by different people to arrive at the same conclusions about students' performance), and
C. Items are appropriate for learners within the target grades who have varied levels of proficiency with the LTs (topics).

In this paper we will provide evidence for the three subclaims and thus for the primary claim about the appropriateness of the CT items. In the discussion section we return to an evaluation of these claims and look ahead to other claims that we would like to evaluate in future work.

Assessment triangle

The assessment triangle is a framework for representing assessment as a process of reasoning from evidence, focused on three interconnected elements for assessment design and interpretation: cognition, observation, and interpretation (Pellegrino et al., 2001). The cognition vertex refers to cognitive theories and models of knowledge representation, processing, and development in the domain (e.g., CT). The observation vertex refers to the design of assessments that will provide evidence of what students know and can do. As such, it is guided by elements of the cognition vertex. The interpretation vertex refers to how one makes reasoned inferences from the collected observations; it can include both rubrics and statistical models. All three vertices should be considered individually, and in combination with each other (Pellegrino et al., 2001). The assessment triangle, as a framework, can be used to frame how one designs assessments and interprets the results of those assessments (e.g., Streveler et al., 2011).

Evidence-Centered Design (ECD)

Evidence-centered design (ECD) is a framework and a design process that considers assessment design across multiple layers of implementation (Mislevy et al., 2003). While the assessment triangle is a theoretical framework, ECD provides steps and processes to ensure that assessment developers focus on all three vertices of the assessment triangle. Although often used for large-scale assessment design, ECD can also be used to design classroom-based assessments. In essence, ECD represents a space of claims about what students know and can do, an articulation of the evidence that students could provide (in what they say or do) that could be used as evidence for these competencies, and the features of assessment items that could elicit such evidence (Pellegrino et al., 2014). The present research and development process used ECD, most prominently by creating *design patterns* that describe families of assessment items (Mislevy & Haertel, 2006). These design patterns are critical, in part because they define the *knowledge, skills, and abilities* (KSAs) that underlie each of the different CT LTs. In essence, the KSAs become the claimed competencies for which one attempts to elicit evidence through the design of assessment items with certain features (Pellegrino et al., 2014).

The Principled Assessment of Computational Thinking (PACT) project has also used ECD to develop CT assessments. Bienkowski and colleagues (2015) developed four design patterns, each representing a different CT practice (e.g., *Design and implement creative solutions and artifacts*). These design patterns were independent of a specific curriculum, programming language, or grade level. Because of this independence, they allow for reuse across different assessment contexts. These design patterns were then used to develop assessments, for instance for HS students in an introductory CS course (Snow et al., 2019). One key distinction between the PACT ECD work and ours is that the PACT group conducted their own domain analysis, whereas our domain analysis was effectively completed through the previous efforts to develop the LTs (Rich et al., 2017, 2018, 2019, 2020). Those LTs were also critical in informing and "jump-starting" the domain modeling process as they directly fed into our design patterns.

Moving from learning trajectories to assessments

The following sections describe how we used a principled approach to work forward from the LTs to develop the assessment items, using the assessment triangle to frame our description.

Cognition vertex

Our cognition vertex is based on the CT LTs (e.g., Figure 1). As mentioned earlier, the LT learning goals include three statements: an understanding (U) goal, a student action (A) goal, and a summary/consensus goal that describe the overall learning goal. We started by aggregating the learning goals within an LT and then used that set of statements to develop a design pattern for the LT. This procedure is elaborated in the next section.

Observation vertex

Design patterns. We used a knowledge representation called a design pattern that assists in specifying key details of the assessment argument, ensuring that developers are explicit in identifying the claims, evidence for those claims, and methods by which one may elicit that evidence (Mislevy & Haertel, 2006). We created design patterns for five of the six CT trajectories: decomposition, sequence, repetition, conditionals, and variables[2]. Each design pattern was then used to develop a family of assessment items that individually and collectively could elicit evidence that students possessed the knowledge, skill, and abilities associated with respective LT.

There are four main categories of design pattern components: description (title, summary, rationale), KSAs (focal and additional), evidence statements (potential observations and potential work products), and item features (characteristic and variable). To illustrate our process for creating the design patterns we describe these four categories and how the LTs informed their development. We use the repetition LT (Figure 1) and the repetition design pattern (appendix) as an example to illustrate the procedure that was used for all five design patterns. Two design pattern features are worth noting. First, connections to the source LT statements are noted with alphanumeric codes (e.g., "3; 3U"). Interested readers can use these codes to trace content in the design pattern to its source in the LT (i.e., Figure 1). Second, the design pattern uses a convention in which occasional text is struck (e.g., "~~forever~~"); this is used when a LT statement included content that was not covered by the curricular lessons (e.g., in the lessons students did not learn about *forever* loops but did learn about *repeat until* and *repeat X times* loops). Strikethrough text was not used to develop assessment items.

The description category of the LT provides a brief summary and rationale for the topic that is represented in the design pattern. We wrote the description components to capture the full

[2] A design pattern for the debugging LT was not developed due to concerns about the difficulty of attempting to assess debugging, especially when using paper-and-pencil assessments.

range of learning goals expressed in the LT, occasionally using language from specific learning goals (e.g., 5A in the repetition design pattern summary).

The knowledge, skills, and abilities (KSAs) category describes the cognitive foundations necessary for successfully responding to items written from this design pattern. Identifying and defining focal KSAs allows one to articulate the multiple proficiencies that form the design pattern and thus measurement targets; these focal KSAs can also be used to develop an ECD student model (Bienkowski et al., 2015; Mislevy et al., 2003). We used CT LT statements to define focal KSAs. We combined, reordered, and collapsed statements when needed, in the process rewording them. For example, the focal KSA "*Knowledge that different kinds of tasks require different kinds of repeated instructions and therefore different repeat commands*" is a combination of LT statements 1, 4.1, 4.1U, 5.1, and 5.1U (see appendix and Figure 1). Additional KSAs represent additional cognitive factors that might be required to respond to items, but are not measurement targets. This step of defining KSAs most clearly illustrates the hypothesized cognitive connections between the assessment items and the LT.

Design patterns also include potential observations and potential work products. Once KSAs are articulated, one can define the potential observations and potential work products that would provide evidence for whether a student can use those KSAs. Collectively the potential work products and potential observations are evidence statements—descriptions of what students could say or do that would count as objective evidence that they have the competencies claimed by the design pattern (Pellegrino et al., 2014). As with the KSAs, we used the LT statements to help write evidence statements, along with considerations about how students could demonstrate the KSAs. Often, the student action goal was useful in writing the potential observations (e.g., 4.1A).

Finally, design patterns include characteristic item features and variable item features. Together, the potential observations and student work products suggest and constrain potential features of assessment items that might elicit the intended evidence. Characteristic item features define the features that all items developed from the design pattern will have. In contrast, a variable item feature defines a feature that some items will have, or a dimension along which to vary items. When creating the design patterns, some of the item features were written based on a LT action goal while others were written from the KSAs. One variable feature present in all five design patterns was the presence of a mathematics application/context. A subsidiary characteristic feature was that such items should use whole numbers, not fractions. In this way, we connected to the curricular lessons, but ensured that those students who were still developing their fraction understanding would not be precluded from demonstrating their CT. We used this approach to try to minimize construct-irrelevant variance (AERA, APA, & NCME, 2014).

In summary, the LT statements were used to write claimed competencies articulated in the form of focal KSAs. These KSAs were used (in concert with other LT statements) to develop the remainder of the design pattern (i.e., evidence statements and task features). These design pattern components were in turn used to guide the development of the assessment items. This allowed us to use the LTs to define the cognition vertex of the assessment triangle, and to use ECD to link the cognition and observation vertices.

Developing Assessment Items. We used the design patterns to develop individual assessment items that might elicit one or more focal KSAs. Potential observations and work products were used to suggest item designs, and the characteristic and variable features were used to suggest how to create and vary items from the same design pattern. In doing so, we used a variety of response formats, including different types of selected and constructed response formats. Some items used the Scratch interface and/or code blocks and others did not. Because design patterns define a family of assessment items, the developed items are not exhaustive; future developers can use the design patterns to develop additional items that align to the LTs.

As part of the item development process all items included an *exemplar student response* that signaled what a student response would look like for a student that had the relevant KSAs. The focal KSAs and evidence statements were used to guide and check the exemplar student responses. Before asking students to complete the assessment items, all items were reviewed by project members with experience teaching CT and writing CT curriculum for the target grade levels. These reviews focused on item clarity, language, accessibility, and alignment to the learning goals. Items were revised following review.

Developing Assessment Instruments. The next step was to assemble individual items into instruments. We created six instruments. Each grade had three instruments that were designed to be administered after specific lessons were enacted (the assessments were administered after completing sets of four lessons and were labeled according to the administration chronology, i.e., Early, Mid, and Late). This allowed us to select items to match the content that was being learned at the corresponding time. Because the content of the lessons constrained the relevant assessment items, a single instrument often had different numbers of items for each LT and included items from only some LTs. Attempting to include assessment items from all LTs in equal number on each instrument would have created misalignment between the assessments and students' opportunities to learn the content.

Each instrument sampled from the content of the different LTs that students had been taught in the preceding curricular lessons, included approximately 10 assessment items, and was designed to be completed in one class period. We selected items in approximate correspondence to the distribution of LT goals throughout the associated lessons. For example, the Grade 3 Early assessment had decomposition items (n = 5), sequence items (n = 3), and repetition items (n = 2), while the Grade 4 Early assessment had decomposition items (n = 3), sequence items (n = 3), and variables items (n = 4). Forty-four items were used across the six instruments. Although some items were used on multiple instruments, the majority of items appeared on only one instrument. This allowed us to maximize the number of items that we could collect data for, while limiting the length of each assessment instrument.

Pilot testing

During the 2018-2019 academic year, we piloted the curriculum and assessment instruments. Partly because this was a pilot year for curriculum and assessment use, data collection was less complete than planned. As the academic year progressed, the number of teachers that used the assessment instruments decreased due to constraints on classroom time. We were able to collect data on five of the six assessment instruments (all except the Grade 4 Late); of these five, the two Early assessments (Grade 3 and Grade 4) were used the most widely. Therefore, in this paper we focus on these two instruments because they have the most student data. The analytic approach we describe provides an initial validity evidence and illustrates the approach that could be used with the other four instruments. As elaborated in the Discussion section, this paper describes only a portion of the larger assessment use and validation approaches that we are engaged in to develop LT-based CT assessments for upper elementary students.

Context and participants

Data were collected from students in five elementary schools (four public schools; one private school) in two midwestern states in the U.S. with diverse student populations. The school-level demographic data for each school is reported in Table 1. For the four public schools where state reports were available, each had a diversity in cultural backgrounds, socioeconomic status (i.e., from low income families or needing free/reduced price meals), educational needs (i.e., enrolled in special education), and language background (i.e., English language learners). Five Grade 3 classes and three Grade 4 classes had students that completed the assessment instruments. We obtained parental consent and assent for students in the participating classes for data collection. The assessment instruments were administered after the teachers (with researcher/curriculum developers' assistance) enacted the integrated lessons. As this was our first year to pilot both the lessons and assessments, students in Grade 4 had not received prior Grade 3 instruction as would be expected in a full implementation of the curricular lessons.

Scoring (interpretation vertex)

First, we determined how to interpret students' responses using objective criteria. We coded student responses for accuracy and characteristics of their responses. Each item's exemplar response was used to signal what a "correct answer" was, or what feature(s) a correct answer would have. The majority of items had simple scoring criteria that corresponded to a correct answer for the selected response items and for some constructed response items. When items resulted in a range of student responses, rubrics were developed to aid scoring. Rubrics were developed after reviewing a subset of student responses and considering the type of knowledge and skills that the assessment items were designed to elicit. This process resulted in different categories of student responses that were then coded using the rubric to determine a single score. An example rubric (for the item shown in Figure 3) is provided in Table 2. This rubric has three categories (scores) that correspond to different levels of a student response that might show evidence of using the repetition construct. Student responses were identified that would typify responses for each category, and added to the rubric as examples. In the first category (score = 0) the student might have illustrated the correct outcome, but without using repetition (e.g., they designate each of the friends should receive three cookies, but they do not use any repetition to do it). In this case, because the assessment item is aligned to the repetition trajectory, it is explicitly aimed at providing students an opportunity to demonstrate they can use repetition; providing an alternate method to achieve the correct outcome, without using repetition, does not provide evidence that that student can write instructions using repetition. The item was designed to elicit this ability (the item instructs students to "Use the instruction 'repeat 3 times' at least one time."). In the second category (score = 1) the student might have shown they understood that an instruction/command could be issued once and then repeated (perhaps multiple times), however, the student's instruction does not achieve the intended outcome. In contrast, in the third category (score = 2), the student both achieves the intended outcome and gives an indication that that an instruction/command is to be repeated.

To ensure reliable coding of students' responses we used a dual-rater method with 100% overlap of dual-coded responses. This method involved an initial training session, followed by coding a subset of the data, an interim check on agreement, and then a complete coding of the data. In the initial training session both coders and the lead author collectively applied the scoring criteria and rubrics to three students' responses. During this session adjustments and clarifications to the scoring criteria or rubrics were made as needed. The two coders then independently scored approximately 30% of the responses. Percent agreement was calculated and disagreements were resolved through discussion between the two coders, moderated by the lead author; particular attention was paid to any items where agreement was below 80%. During these discussions we revised the scoring criteria or rubrics as needed to clarify scoring decisions. Following this interim check the two coders scored the remaining 70% of the data. Again percent agreement was calculated and disagreements were resolved through discussion; scoring criteria and rubrics were updated as needed to reflect the scoring decisions that were used or finalized during that final session.

We report reliability using Cohen's Kappa (weighted) which can range between -1 and 1, with a higher value indicating greater agreement among raters. Weighted Kappa allows one to compute agreement among ordinal scores while allowing scores that are closer to yield higher agreement (Viera & Garrett, 2005). We computed Cohen's Kappa for each item (excluding the three students' responses that were used for each training session), for each assessment instrument. Overall, we had high reliability on both the Grade 3 Early instrument (median = .98, min = .94, max = 1) and the Grade 4 Early instrument (median = .91, min = .55, max = 1).

Results and findings

We conducted a multi-step analysis of the student response data, analyzing each grade's instrument separately. When reporting these results we reference the item codes to identify items. Our item code naming convention denotes the trajectory (e.g., "DC" = decomposition, "S" = sequence, etc.) and the item number (e.g., "01", "02", etc.), and uses a final indicator to distinguish between variations in similar items (e.g., "a", "b", etc.).

Grade 3 Early assessment

Ninety students from five classes (four teachers; one teacher taught two classes) completed the Grade 3 assessment following instruction. Before data analysis, we dropped one of the decomposition items (DC.04.a) because of concerns raised by project members that it was not measuring any of the decomposition KSAs. The remaining nine items sampled from items developed to align to one of three design patterns/LTs: decomposition, sequence, and repetition. After initial data exploration, we dichotomized the three items (DC.08.a, S.01.a, R.01.a) that had been coded polytomously (i.e., using partial-credit scoring as illustrated by the rubric in Table 2). For each item scored polytomously, we collapsed coding across score categories based on the item, rubric levels, and distribution of student responses among rubric levels. For instance, continuing with the R.01.a example (Figure 3 and Table 2), we collapsed two categories (where score = 1 or score = 2) into a single category (score = 1) and retained the remaining category (where score = 0).

To understand how the items were performing individually, we conducted classical test theory (CTT) analysis. Difficulty (p) values indicate the "easiness" of an item and so (counterintuitively) higher difficulty values indicate easier items and lower values indicate harder items (Crocker, 2006). For dichotomous items, difficulty values are the proportion of correct responses. Item R.01.a was the most challenging item for our sample of students to answer. Item DC.03.b, a multiple-choice item about why one should decompose an item, was the easiest item on the instrument with almost 80% of students endorsing a correct answer. All item difficulty values fell within the .20–.80 range of acceptability (see Table 3). We also calculated discrimination, a measure of how well an item can discriminate between students of "high" and "low" proficiency on the construct of interest. A higher discrimination value indicates that the item is better able to discriminate between these levels; a low (or negative) discrimination value indicates there might be problems with the item. All items had discrimination values greater than .20, a minimum value typically used as a cut-off to signal "acceptable" items (see Table 3).

To further understand these data, we conducted analyses using Item Response Theory (IRT) using the eRm package (Mair & Hatzinger, 2007) in R. IRT (Baker, 2001; Embretson & Reise, 2000) allows one to model students and items on the same latent trait (θ), enabling one to consider both item and student (person) characteristics simultaneously (e.g., Figure 4). Although a small sample, we fit a Rasch model to the data. An overall goodness of fit test (Anderson LR-test, $X^2(6) = 11.14$, $p = .08$), indicated the Rasch model fit the data. Figure 4 plots both person and item characteristics, notably ability (persons) and difficulty (items). The top quarter of the plot shows a histogram of person ability as estimated for students in our sample, where the histogram bars indicate the count and location on the latent trait scale. The remainder of the plot shows items' locations on the latent trait scale. Each point indicates where the item falls relative to a 50% probability that students would answer the item correctly as predicted on the basis of their overall ability estimate (based on their performance on the instrument as a whole). Item difficulty (β) showed a good range of measurement along the latent trait, with values ranging from -1.18 to +2.23 (see Figure 4). Although there were a small number of items, the items show a range of difficulties, a desirable feature for an assessment that will be used with students with a range of proficiencies on the CT construct.

Overall, students' performance on the assessment instrument was moderate. Students answered more than half of the nine items correctly ($M = 5.5$, $SD = 2.5$). As indicated earlier, this is a desirable finding for collecting initial evidence of the appropriateness of the instrument for use with students that have a varying range of abilities. For potential diagnostic and formative use one does not want a test that is too easy or too hard because it can limit the utility of the information provided by the assessment. Further, considering these items in light of the LTs from which they were designed, it appears that the decomposition items used on this instrument were the easiest items, followed by the sequence items, followed by the repetition items (as indicated by the difficulty estimates provided by both the CTT and IRT analysis).

Grade 4 early assessment

Fifty-four students from three classes (three teachers) completed the Grade 4 Early assessment following instruction. Before analysis we dropped the DC.04.a item for the same reason we dropped it from the Grade 3 assessment. We also dropped a variables item (V.07.c) because of a typographical error that created ambiguity around what students were expected to do to answer the item correctly. After dropping Items DC.04.a and V.07.c, eight items remained. These items sampled from the variables, sequence, and decomposition LTs. After initial data exploration, we recoded one item (S.02.a) from polytomous to dichotomous. Additional data exploration involved looking at measures of internal consistency and correlations among items and total score. During this exploration we identified four potentially problematic items and examined the items and scoring criteria to understand whether the items might need to be removed before further data analysis. This investigation did signal problems with two of those four items. Item V.10.a used a multiple-choice format where students were asked to circle multiple correct answers (this format was potentially difficult because the instructions about selecting multiple responses were unclear). Perhaps more significant though, the content that the item was designed to assess was not reflected well by the content in the lessons. Item S.06.b used a true/false format. Upon review we determined that it used a statement that—although true in most cases—could reasonably be argued as false by someone experienced with Scratch. Because Items V.10.a and S.06.b were negatively correlated with total score and their review identified potential problems, we decided to drop them before conducting further analysis.

Difficulty and discrimination values for the six items that remained are reported in Table 4. Discrimination values were all acceptable. Three items had difficulty values greater than .80, indicating a large proportion of students got them correct and that they were easy for most students in our sample. This finding is further explored in the IRT results when we compare the items to students' ability estimates.

Although we had only a small sample of student responses, we fit a Rasch model to the data (Anderson LR-test $X^2(3) = 7.09$, $p = .07$). In calculating overall model fit two items had to be excluded due to their student response patterns (which is in part due to their easiness which resulted in little variance among subgroups). IRT difficulty showed a good range of measurement along the latent trait, with values ranging from -1.48 to +1.54. As with the Grade 3 Early instrument, items were spaced out along the latent dimension, indicating that some items (e.g., DC.02.b and V.04c) can be answered correctly even by students that are estimated to have lower CT proficiency, and some are only likely to be answered correctly by those that are estimated to have a higher CT proficiency (V.03.b and S.02.a) (see Figure 5).

As with the Grade 3 Early instrument, students' performance on the Grade 4 Early assessment was moderate. On average, students answered more than half of the six items correctly ($M = 3.8$, $SD = 1.4$). Unlike the results for the Grade 3 Early instrument, there did not seem to be any clustering of item difficulty by LT, although both decomposition items were among the easier items on the assessment. With this sample, the two sequence items were the most highly discriminating (see Table 4).

Discussion

In this paper, we have shown how the three vertices of the assessment triangle were defined and coordinated, using ECD, to yield specific claims regarding the validity of CT assessments for use with upper elementary students. Our development methodology constitutes evidence for one part of our validity argument (Crocker, 2006; Pellegrino et al., 2016), and the empirical results constitute additional evidence for other parts of the validity argument.

Assessment development

Validation requires more than the application of psychometric procedures (Douglas & Purzer, 2015). It requires a principled design process that documents design decisions and collects multiple sources of evidence to support the proposed interpretation and use of assessment results (Pellegrino, 2013; Gane et al., 2018). We presented initial, positive evidence to demonstrate that one can use LTs to design assessments in a principled way and use those assessments to collect evidence that students can draw upon their CT knowledge, skills, and abilities to respond appropriately to the assessment tasks following instruction. The LTs provided a cognitive model, crucial for implementing the other parts of the assessment triangle (the observation and interpretation vertices), as well as for guiding the assembly of evidence regarding validity. These key aspects of design and validation have been notably absent from many prior attempts to develop CT assessments (Tang et al., 2020).

By explicating our assessment development method, we traced how we worked forward from the LTs, to design patterns, to assessment items, and finally to assessment instruments. In doing so, we highlighted critical details that are needed for a comprehensive validity argument.

Empirical results

Since this was our first round of implementation and data collection, it constitutes a pilot evaluation of both the lessons and the assessments. The number of participating students and teachers was not large, and this paper only reports on two of the six assessment instruments that were developed. As such, there are limitations on the interpretations and claims we can make given the empirical data. Despite those constraints, we were able to show that the two early assessment instruments and associated items had acceptable levels of interrater reliability. Our analysis of item and test performance signaled a small number of items that needed revision or exclusion, which we completed before more recent data collection efforts.

For the majority of items, CTT and IRT analyses provided evidence that the items were appropriate for students with a range of CT proficiencies. The items showed a range of difficulties, and acceptable levels of discrimination. Students performed reasonably well on the items as a whole, consistent with the conclusion that students have the knowledge and skills needed to answer these questions after instruction with the lessons. If the items had been too difficult for students, they would have appeared further to the right in the Person-Item maps in Figures 4 and 5 (e.g., $\theta > 2$). Although these performance data suggest students learned from the lessons, we do not have the counterfactual (students with no CT instruction). Therefore, we do not know how students would have performed if they had not completed the curricular lessons.

Reflecting on the assessment item claims and evidence for these claims

As stated previously, we have one primary claim that we support in this paper: the CT assessment items measure students' performance with respect to the LTs (topics) that students were introduced to during the lessons. This primary claim involves three constituent subclaims, for which we have provided evidence in this paper. **Subclaim A (items are aligned to the LTs and students' opportunities to learn)** was supported by the evidence we have provided related to our process and our documentation of this process during development. In particular, we used a principled approach to convert the LTs into design patterns that were then used to write assessment items. These assessment items were reviewed by individual project members with experience teaching CT and writing CT curriculum for the target grade levels and then revised following review. Finally, assembling the items into instruments was done using the curriculum developers' mapping between lessons and LT goals to ensure that the content used to design the lesson activities was represented in the items that appeared on the assessment instruments, and that the items on the instrument did not go beyond the content used in the lessons. In this way, we were able to align items to both the LTs and students' opportunities to learn. **Subclaim B (item scores are reliable)** is supported by the evidence we have provided that the item scoring rules and rubrics can be used by different people to arrive at the same conclusions about students' performance. In particular, the interrater reliability results showed high agreement among scorers for all items and across both instruments. In addition, the process we used to develop and modify the scoring rules and rubrics during the training and subsequent use add to the evidence supporting our claim that the scores can be reliably determined. **Subclaim C (items are appropriate for learners within the target grades who have varied levels of proficiency with the LTs/topics)** is supported by the empirical evidence on the difficulty and discrimination of the items, along with the comparison between item difficulty and student proficiency (i.e., Figures 4 and 5).

Although we are still engaged in assessment development, we present initial evidence for our primary claim and subclaims. By evaluating the primary claim we also provide support for the cognitive and inferential components of validity (Pellegrino et al., 2016). Subclaim A is most related to the cognitive component of validity. Evidence for Subclaim A applies to all the items that were created from the LT design patterns, because the same process was used for items across the five LTs and six instruments; the design evidence applies to even items that were not piloted. Subclaims B & C are most related to the inferential components of validity. For Subclaim B and C, the evidence we have in support of these subclaims is limited to those items used on the Grade 3 and Grade 4 Early instruments. Eventually, our goal is to provide evidence in support for these two subclaims for all items across the five LTs and six instruments.

Future directions

Validation is an ongoing process (Douglas & Purzer, 2015) that requires accumulating multiple sources of evidence tailored to the specific claims being advanced (AERA, APA, & NCME, 2014; Pellegrino et al., 2016). Doing so requires multiple methodologies, sources of data, and types of analysis. As such, this paper presents an initial assembly of validity evidence and is not the full extent of our efforts to validate and use these assessments. For instance, we have also conducted cognitive interviews with students to further understand the knowledge and skills elicited by the assessment items relative to the targeted components of the specific LTs (Luo, 2020; Luo et al., 2020). In brief, these interviews helped identify items that needed revisions (e.g., to minimize the possibility that math knowledge or non-targeted CT knowledge could interfere with students' ability to respond to an item) and provided evidence that specific CT assessment items seemed to elicit the relevant KSAs in the specific LT.

We have revised the assessments following our 2018-2019 pilot implementation, including editing individual assessment items, the overall instruments, and the design patterns. Such iterative design is critical in developing and validating assessments. We are currently analyzing data collected from the use of these revised assessments during AY 2019-2020. The analytic and empiric approach we describe in this paper can be used for the other items with these new data, including items that appeared on the other four assessment instruments (i.e., the Mid and Late instruments in Grades 3 and 4).

Because this 2019-2020 data collection involved a greater number of teachers and students, we are hopeful that these larger sample sizes will enable more sophisticated psychometric modeling, potentially including structural analysis (e.g., to determine dimensionality and investigate the use of sub-scores), further IRT analysis (using 2PL models), and diagnostic classification modelling (Pellegrino et al., 2014). Such data would allow us to advance and support additional claims about the appropriateness of the instruments (i.e., the specific assemblages of items), the utility and appropriateness of calculating sub-scores tied to individual LTs, and the use of these overall and/or sub-scores to understand student performance and growth over time. However, we also know that further validation efforts must be carefully considered in light of the claims being made for the assessments and uses. As Douglas and Purzer (2015) point out, it can be tempting to equate the indiscriminate application of psychometric approaches with validation, but doing so is a mistake. Our design of the CT items and instruments allows for various psychometric analyses (e.g., factor analysis, 2PL IRT, diagnostic classification modeling) that could be informative and support interesting claims. However, expectations for engaging in these efforts must be tempered with a dose of reality. The limiting factor is the number of student responses needed to perform these analyses. For data to be meaningful, students must have had the appropriate opportunities to learn. Therefore, data collection is limited to teachers and students that are using the curricular lessons, and faces the realities of limited time and student attention. Elementary students will not sit for long assessments, and teachers are reluctant to give up instructional time for assessments that do not translate into grades for students. These are real and valid constraints on data collection and limit the potential for using sophisticated psychometric approaches (more common in large scale testing) with data collected from classroom assessments completed as part of students' in-class activities.

In addition to our efforts to further validate the assessments, we are also interested in exploring different uses of the assessments, particularly in ways that can support teachers' instruction and students' learning. Because we used a design method that entails documenting the rationale underlying design decisions and item alignment, these assessments do not have to be used as "intact" instruments. Instead they could be disassembled and reassembled based on the specific lessons that are taught or the specific content that one wants to assess. For instance, one could assemble items on the basis of the LT from which they were developed in order to create a progress variable against which students' performance could be measured and tracked (Pellegrino et al., 2016). In the future, these assessments might be used by teachers as diagnostic instruments that could be used formatively, summatively, and/or in a pre–post design. Such uses are enabled by the design process but would also require additional validation. To date we have focused on using a principled design process to develop the assessment items and associated scoring criteria/rubrics, and on collecting initial evidence to support the claim that the CT assessment items measure students' performance with respect to the LTs (topics) that students were introduced to during the lessons. How to support teachers in using these assessments formatively represents future work.

Final thoughts

As the field seeks to make advances in CT assessment, attention must be paid to the connections between learning and assessment, especially as they are mediated by processes of curriculum design and instruction. As argued in *Knowing What Students Know: The Science and Design of Educational Assessment* (Pellegrino et al., 2001), we need to strive for coordination among curriculum, instruction, and assessment. Such coordination is only possible when all three are guided by theories, models, and data about learning and knowing in the domain of interest. We have adhered to this conceptualization in our initial work on the design of CT assessments. Our design process allowed us to create assessment items and organize them into multiple instruments, mapped to different phases of instruction, where the content assessed was based on the same learning goals (LTs) that were used to design the curricular lessons. Although some items were identified for revision, the majority performed well. The process of using the LTs to develop assessments tied to curricular lessons is one of several possible approaches to the assessment of CT. Rather than trying to provide a general evaluation of students' CT, our assessments were designed to cover multiple facets of CT (i.e., the different LTs) that students would be expected to learn through instruction based on a set of carefully designed lessons. Furthermore, the assessments were not developed directly from the content of lessons but focused on the CT constructs embedded in the lessons. As such, they are best considered proximal assessments (Ruiz-Primo et al., 2002), and are appropriate for measuring the effect of the curriculum and instruction while also serving as a measure of near transfer (Pellegrino et al., 2014).

Acknowledgments. This work was supported by the [BLIND] under Grant [BLIND]. We would also like to acknowledge the assistance of [BLIND] in reviewing assessments and design patterns, [BLIND] in facilitating data collection, and [BLIND] in scoring student responses. A previous version of this paper presenting preliminary design and empirical results appeared in the ICLS 2020 conference proceedings.

References

AERA, APA, & NCME. (2014). Standards for educational and psychological testing (2014 ed.). Washington, DC: American Educational Research Association.

Barr, D., Harrison, J., & Conery, L. (2011). Computational thinking: A digital age skill for everyone. *Learning & Leading with Technology*, *38*(6), 20-23.

Barr, V., & Stephenson, C. (2011). Bringing computational thinking to K-12: what is Involved and what is the role of the computer science education community?. *ACM Inroads*, *2*(1), 48-54.

Basawapatna, A., Koh, K. H., Repenning, A., Webb, D. C., & Marshall, K. S. (2011). Recognizing computational thinking patterns. In *Proceedings of the 42nd ACM Technical Symposium on Computer Science Education* (pp. 245-250).

Bers, M. U., Flannery, L., Kazakoff, E. R., & Sullivan, A. (2014). Computational thinking and tinkering: Exploration of an early childhood robotics curriculum. *Computers & Education*, *72*, 145-157.

Bienkowski, M., Snow, E., Rutstein, D., & Grover, S. (2015). *Assessment design patterns for computational thinking practices in secondary computer science: A first look.* (SRI technical report). Menlo Park, CA: SRI International. Retrieved from http://pact.sri.com/resources.html

Boyer, T. W., & Levine, S. C. (2012). Child proportional scaling: Is 1/3= 2/6= 3/9= 4/12?. *Journal of Experimental Child Psychology*, *111*(3), 516-533.

Brennan, K., & Resnick, M. (2012). New frameworks for studying and assessing the development of computational thinking. Paper presented at the *2012 Annual Meeting of the American Educational Research Association*, Vancouver, Canada.

Clements, D. H., Sarama, J., Baroody, A. J., & Joswick, C. (2020). Efficacy of a learning trajectory approach compared to a teach-to-target approach for addition and subtraction. *ZDM*, 1-12.

Computer Science Teachers Association [CSTA]. (2017). CSTA K-12 computer science standards, revised 2017. *Computer Science Teachers Association,* USA.

Crocker, L. (2006). Introduction to Measurement Theory. In J. Green L., G. Camilli, & P.B. Elmore (Eds.), *Handbook of complementary methods in education research*. Mahwah, NJ: Erlbaum.

Douglas, K. A., & Purzer, Ş. (2015). Validity: Meaning and relevancy in assessment for engineering education research. *Journal of Engineering Education, 104*(2), 108–118.

Douglas, K., A., Gane, B. D., Neumann, K., & Pellegrino, J. W., (2020). Contemporary methods of assessing integrated STEM competencies. In C. C. Johnson, M. Mohr-Schroeder, T. Moore, L. Bryan, & L. English (Eds.) *Handbook of research on STEM education*. New York, NY: Routledge/Taylor & Francis.

English, L. D. (2017). Advancing elementary and middle school STEM education. *International Journal of Science and Mathematics Education, 15*(1), 5-24.

Ericson, B. J., Margulieux, L. E., & Rick, J. (2017). Solving parsons problems versus fixing and writing code. In *Proceedings of the 17th Koli Calling International Conference on Computing Education Research* (pp. 20-29).

Gane, B.D., Zaidi, S.Z., & Pellegrino, J.W. (2018). Measuring what matters: Using technology to assess multidimensional learning. *European Journal of Education, 53*, 176–187. https://doi.org/10.1111/ejed.12269

Harwell, M., Moreno, M., Phillips, A., Guzey, S. S., Moore, T. J., & Roehrig, G. H. (2015). A study of STEM assessments in engineering, science, and mathematics for elementary and middle school students. *School Science and Mathematics*, *115*(2), 66-74.

Israel, M., Pearson, J. N., Tapia, T., Wherfel, Q. M., & Reese, G. (2015). Supporting all learners in school-wide computational thinking: A cross-case qualitative analysis. *Computers & Education*, *82*, 263-279.

K-12 Computer Science Framework (2016). *K-12 computer science framework.* https://www.k12cs.org/

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*, 527–535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.) *Educational Measurement* (4th Ed., pp. 17-64). Westport, CT: Praeger Publishers.

Lee, I., Martin, F., & Apone, K. (2014). Integrating computational thinking across the K--8 curriculum. *Acm Inroads, 5*(4), 64-71.

Luo, F. (2020). *Exploring elementary students' computational thinking leveraging affordances of learning trajectories* [Unpublished doctoral dissertation]. University of Florida.

Luo, F., Israel, M., Liu, R., Yan, W., Gane, B., & Hampton, J. (2020). Understanding students' computational thinking through cognitive interviews: A learning trajectory-based analysis. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education (SIGCSE '20)*. New York, NY: Association for Computing Machinery.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*(9).

Maloney, A. P., Confrey, J., & Nguyen, K. H. (Eds.). (2014). *Learning over time: Learning trajectories in mathematics education*. IAP.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research & Perspective, 1*(1), 3–62.

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

Moreno-León, J., Robles, G., & Román-González, M. (2015). Dr. Scratch: Automatic analysis of scratch projects to assess and foster computational thinking. *RED. Revista de Educación a Distancia*, (46), 1-23.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: The National Academies Press.

Pellegrino, J. W., DiBello, L. V., & Brophy, S. P. (2014). The science and design of assessment in engineering education. In A. Johri & B. M. Olds (Eds.), *Cambridge Handbook of Engineering Education Research* (pp. 571–598). Cambridge: Cambridge University Press.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59–81.

Price, T. W., & Barnes, T. (2015). Comparing textual and block interfaces in a novice programming environment. In *Proceedings of the eleventh annual international conference on international computing education research* (pp. 91-99).

Resnick, M., Maloney, J., Monroy-Hernández, A., Rusk, N., Eastmond, E., Brennan, K., ... & Kafai, Y. (2009). Scratch: programming for all. *Communications of the ACM, 52*(11), 60-67.

Rich, K. M., Strickland, C., Binkowski, T. A., Moran, C., & Franklin, D., (2017). K-8 learning trajectories derived from research literature: Sequence, repetition, conditionals. In *Proceedings of the 2017 ACM Conference on International Computing Education Research (ICER '17)*. New York, NY: ACM.

Rich, K. M., Binkowski, T. A., Strickland, C., & Franklin, D. (2018). Decomposition: A k-8 computational thinking learning trajectory. In *Proceedings of the 2018 ACM Conference on International Computing Education Research* (pp. 124-132).

Rich, K. M., Strickland, C., Binkowski, T. A., & Franklin, D. (2019). A k-8 debugging learning trajectory derived from research literature. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (pp. 745-751).

Rich, K., Franklin, D., Strickland, C., Isaacs, A., & Eatinger, D. (2020). *A learning trajectory for variables: Using levels of thinking to develop instruction for particular contexts*. [Manuscript submitted for publication].

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.

Román-González, M., Moreno-León, J. & Robles, G. (2019). Combining assessment tools for a comprehensive evaluation of computational thinking interventions. In *Computational Thinking Education* (pp.79-98). Institute for Educational Science, Paderborn University: Paderborn, Germany.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39*(5), 369–393.

Santo, R., DeLyser, L. A., Ahn, J., Pellicone, A., Aguiar, J., & Wortel-London, S. (2019). Equity in the who, how and what of computer science education: K12 school district conceptualizations of equity in 'CS for all' initiatives. In *2019 Research on Equity and Sustained Participation in Engineering, Computing, and Technology (RESPECT)* (pp. 1-8). IEEE.

Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review, 22*, 142-158.

Seiter, L., & Foreman, B. (2013). Modeling the learning progressions of computational thinking of primary grade students. In *Proceedings of the ninth annual international ACM conference on International computing education research* (pp. 59-66).

Snow, E., Rutstein, D., Basu, S., Bienkowski, M., & Everson, H. T. (2019). Leveraging evidence-centered design to develop assessments of computational thinking practices. *International Journal of Testing, 19*(2), 103-127.

Streveler, R. A., Miller, R. L., Santiago-Roman, A. I., Nelson, M. A., Geist, M. R., & Olds, B. M. (2011). Rigorous methodology for concept inventory development: Using the 'assessment triangle' to develop and test the Thermal and Transport Science Concept Inventory (TTCI). *International Journal of Engineering Education, 27*(5), 17.

Strickland, C., Rich, K. M., Eatinger, D., Lash, T., Isaacs, A., Israel, M., & Franklin, D. (2021). Action Fractions: The design and pilot of an integrated math+CS elementary curriculum based on learning trajectories. In *Proceedings of the 2021 ACM SIGCSE Technical Symposium on Computer Science Education*. ACM.

Tang, X., Yin, Y., Lin, Q., Hadad, R., & Zhai, X. (2020). Assessing computational thinking: A systematic review of empirical studies. *Computers & Education*, *148*, 103798.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine, 37*(5), 360-363.

Voogt, J., Fisser, P., Good, J., Mishra, P., & Yadav, A. (2015). Computational thinking in compulsory education: Towards an agenda for research and practice. *Education and Information Technologies*, *20*(4), 715-728.

Voskoglou, M. G., & Buckley, S. (2012). Problem solving and computational thinking in a learning environment. *Egyptian Computer Science Journal, 36*(4), 28-46.

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology, 25*(1), 127-147.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33-35.

Wing, J. M. (2014). Computational thinking benefits society. *40th Anniversary Blog of Social Issues in Computing*, *2014*.

Yadav, A., Hong, H., & Stephenson, C. (2016). Computational thinking for all: pedagogical approaches to embedding 21st century problem solving in K-12 classrooms. *TechTrends, 60*(6), 565-568.

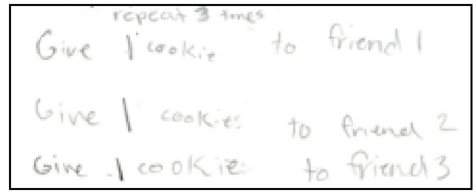
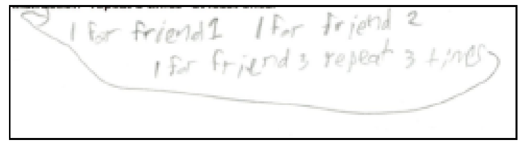
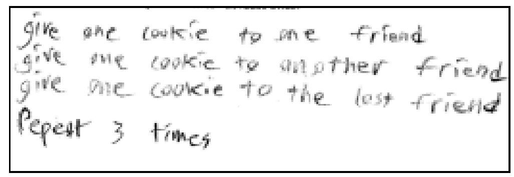
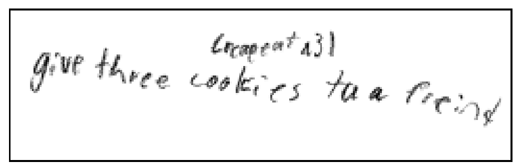
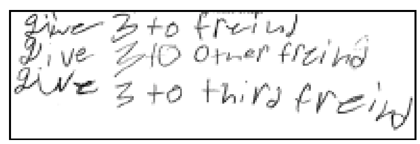
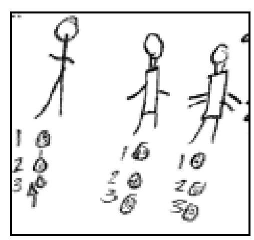
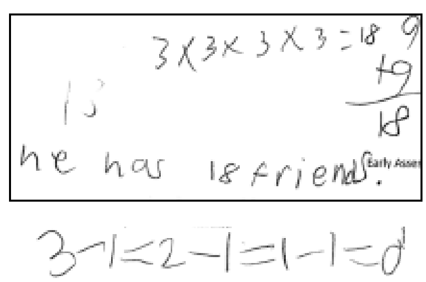
Zhang, L., & Nouri, J. (2019). A systematic review of learning computational thinking through Scratch in K-9. *Computers & Education, 141*, 103607.

Table 1. Demographic data for participating schools (2018-2019 academic year).

School	Ethnicity	Students with IEPs (%)	Low Income Students (%)	English Learners (%)	Grade 3 Classes	Grade 4 Classes
A	40.2% White, 30.9% Black, 9.8% Asian, 6.2% Hispanic, 12.9% Two or More Races	19	59	7	0	1
B	50.6% White, 19.9% Black, 13.7% Asian, 7.9% Hispanic, 7.9% Two or More Races	13	43	13	1	0
C	51.2% White, 20.4% Black, 11.2% Asian, 7.2% Hispanic, 10.1% Two or More Races	13	31	9	1	0
D	62.1% White, 28.2% Hispanic, 2.9% Black, 2.4% Asian, 0.2% American Indian, 4.2% Multicultural	19	43	19	1	1
E	[no data]	[no data]	[no data]	[no data]	2	1

Source: Illinois Report Card (https://www.illinoisreportcard.com/) & Indiana Department of Education.

Table 2. Example rubric for assessment task (R.01.a).

Score	Description	Examples of student responses
2	Instructions use the command "repeat 3 times" and produces the intended outcome.*	 
1	Shows understanding of repetition to get desired result through word explanation or drawings, but instructions won't produce the intended outcome.*	 
0	Incorrect use of repetition concept, or no demonstration of repetition concept to achieve the intended outcome.	  

*Responses with drawings must demonstrate repetition to achieve the intended outcome

Table 3. Item parameters for the Grade 3 Early assessment instrument (after dropping Item DC.04.a).

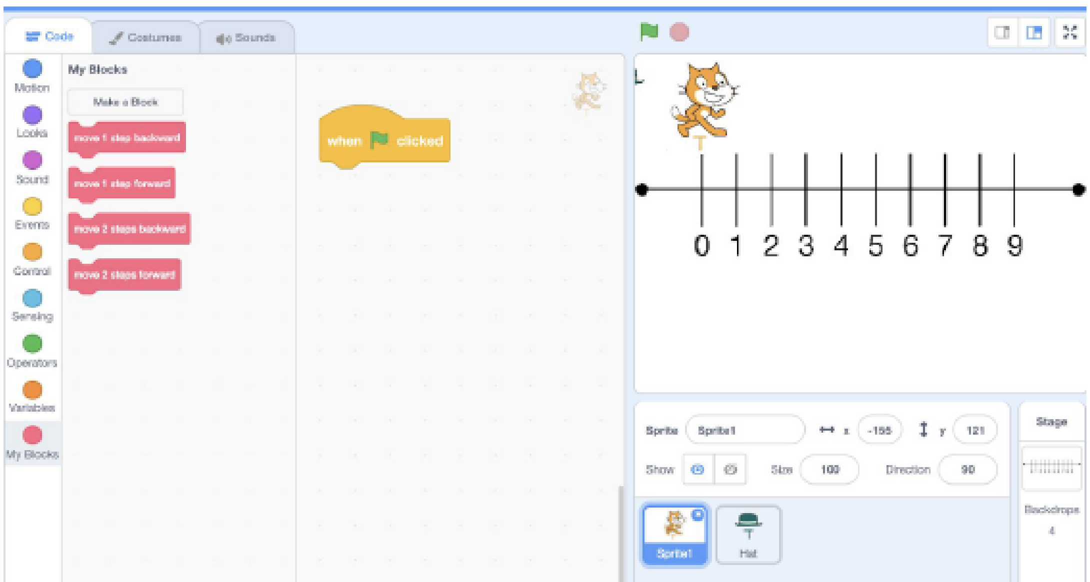
Item	Difficulty (Mean)	Variance	Discrimination
DC.02.a	.72	.45	.47
DC.03.b	.79	.41	.40
DC.05.a	.74	.44	.40
DC.08.a	.62	.49	.63
S.01.a	.52	.50	.87
S.04.b	.60	.49	.87
S.06.a	.66	.48	.47
R.01.a	.24	.43	.53
R.05.a	.57	.50	.90

Table 4. Item parameters for the Grade 4 Early assessment instrument (after dropping Items DC.04.a, V.07.c, V.10.a, and S.06.b).

Item	Difficulty (Mean)	Variance	Discrimination
V.03.b	.35	.48	.44
V.04.c	.85	.36	.22
S.02.a	.31	.47	.77
S.04.d	.57	.50	.72
DC.02.b	.87	.34	.33
DC.03.b	.81	.39	.39

Figure 1. Example learning trajectory (LT): A portion of the full repetition LT, showing the statements that were used to develop the repetition design pattern. Green goals indicate beginning goals and brown goals indicate intermediate, as described by Rich and colleagues (2017); advanced goals are not included in this example. Each set of statements includes a consensus goal, an understanding goal (identified with an "U") and an action goal (identified with an "A").

Figure 2. An example of an assessment item [S.01.a] that uses the Scratch interface and is aligned to the sequence LT. This version includes the exemplar responses which students would not see.



Using the blocks shown, create 2 different scripts (sets of instructions) to move the cat so that he stops at 5 on the number line. Write (or draw) your scripts in the boxes below.

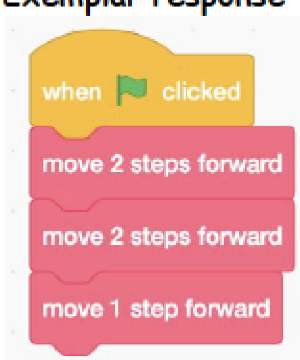
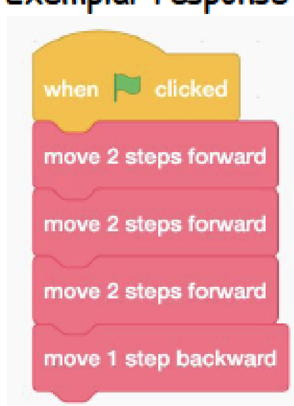
<table>
<tr><td>Script A</td><td>Script B</td></tr>
<tr><td>Exemplar response*:
</td><td>Exemplar response*:
</td></tr>
<tr><td colspan="2">*Any combination of forward and backward steps is acceptable as long as the cat stops at 5.</td></tr>
</table>

Figure 3. An example of an assessment item [R.01.a] that does not use the Scratch interface and is aligned to the repetition LT. This version includes the exemplar responses which students would not see.

Andre has 9 cookies that he wants to give away to 3 friends. He wants to give each friend an equal number of cookies. Write instructions for giving the cookies to the 3 friends. Make sure that each friend gets the same number of cookies. Use the instruction "repeat 3 times" at least one time.

Write your instructions below:

Exemplar response:

1. Repeat 3 times:
    a. Give first friend 1 cookie
    b. Give second friend 1 cookie
    c. Give third friend 1 cookie

[OR]

1. Repeat 3 times:
    a. Give first friend 1 cookie
2. Repeat 3 times:
    a. Give second friend 1 cookie
3. Repeat 3 times:
    a. Give third friend 1 cookie

Figure 4. Person-Item map (akin to a Wright map) that shows the joint distribution of item difficulties (thresholds; lower panel) and person ability estimates (upper panel) for the Grade 3 Early assessment instrument. The upper panel displays a histogram of student ability estimates. The lower panel contains a row for each item, with its associated difficulty indicated with a point.

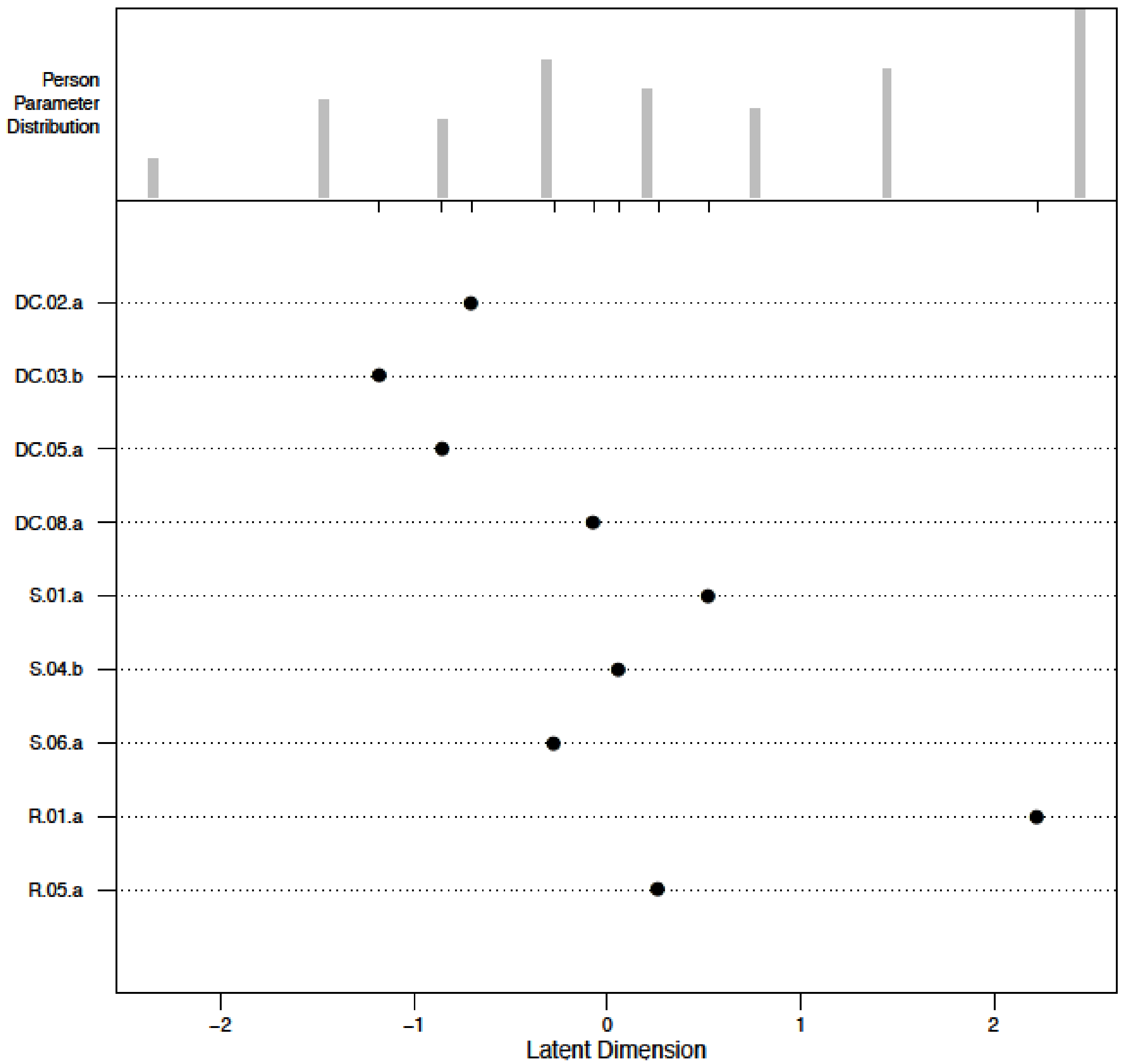
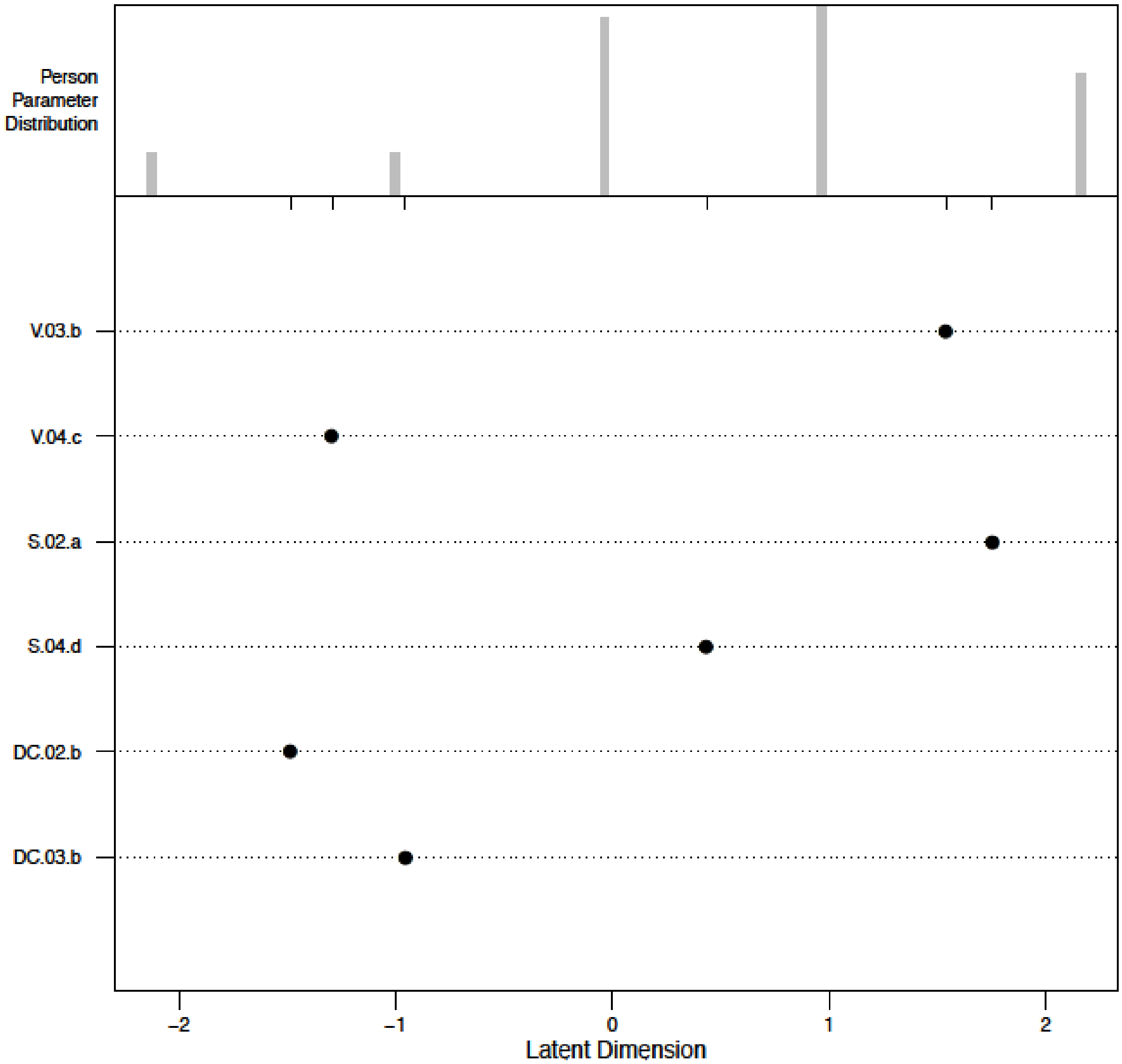


Figure 5. Person-Item map for the Grade 4 Early assessment instrument.



Appendix

Design Pattern: Repetition

<table>
<tr><th>Title</th></tr>
<tr><td>Students who demonstrate understanding can create computational artifacts using repeat language/commands and/or develop instructions/routines using repeat language.</td></tr>
<tr><th>Summary</th></tr>
<tr><td>In this design pattern, a student demonstrates repetition while developing instructions/routines and/or computational artifacts. Using repetition may involve two different approaches: (1) using repeat commands and (2) using the same commands multiple times. Using repeat commands requires the knowledge and ability to select the appropriate repeat command to successfully complete a task (5A). Repeat commands in Scratch include ~~forever,~~ repeat until, and repeat X times.</td></tr>
<tr><th>Rationale</th></tr>
<tr><td>A key aspect of developing efficient computational artifacts is the ability to use repetition effectively. Repetition is a fundamental principle in computational thinking, allowing the same action/code (instructions) to be used multiple times to accomplish an intended goal, while keeping code brief and easy to interpret and reducing the possibility of errors.</td></tr>
<tr><th>Focal Knowledge, Skills, and Abilities (KSAs)</th></tr>
<tr><td>

Knowledge that creating computational artifacts can involve using repeated actions/instructions to accomplish an intended goal (2; 2U)
Ability to Use / Create / Modify a set of instructions that produce the same outcome by using two different approaches: (1) using repeat commands and (2) using the same commands multiple times (3; 3U)
Knowledge that repeat commands (e.g., repeat X times, repeat until~~, forever~~) tell a computer to repeat specific actions/instructions (5; 5U)
Knowledge that different kinds of tasks require different kinds of repeated instructions and therefore different repeat commands (1; 4.1; 4.1U; 5.1; 5.1U)
Ability to use repeat commands (~~forever,~~ repeat until, repeat X times) to Use / Modify / Create instructions that create cumulative effects (1U; 5A)

~~Ability to use / modify / create instructions that use the forever command to accomplish a task that requires infinite repetition~~
Ability to use / modify / create instructions that use the repeat until command to accomplish a task that requires conditional repetition
Ability to use / modify / create instructions that use the repeat X times command to accomplish a task that requires countable repetition

</td></tr>
<tr><th>Additional Knowledge, Skills, and Abilities</th></tr>
<tr><td>

Ability to Create / Use / Modify an ordered set of instructions to produce the intended result [from the Sequence Design Pattern]
Knowledge of (and ability to use) variables [required for repeat X times] and variables and conditionals [required for repeat until]

</td></tr>
<tr><th>Potential Observations</th></tr>
<tr><td>

~~Determine which repeated actions go on forever and which repeated actions stop. (4U)~~

For repeated actions that stop, describe when it will stop. (4A)

</td></tr>
</table>

- For repeated actions that stop, describe the cumulative effect on the outcome, if any. (1A) • Rewrite a set of instructions to replace language that is repeated with repeat language, in order to produce the same outcome. (3A) • Rewrite a set of instructions to replace repeat language with language that is repeated, in order to produce the same outcome. (inverse of 3A) • Identify the type of repetition (~~*forever*,~~ *repeat until*, *repeat X times*) needed to complete a task. (4.1A) • Use the appropriate [repeat] command to create programs that accomplish a task involving repetition. (5.1A) • Predict the cumulative effect of repeating an action (with the repetition communicated using either repeat language or repeating language). • Explain the conditions under which different repeat commands should be used.
Potential Work Products
• A set of instructions (script or pseudocode) that use repetition to accomplish an intended goal • A program that accomplishes an intended goal by using the appropriate repeat command(s) • A statement identifying the result of using repeated actions to accomplish some goal (i.e., a prediction of the cumulative effect of repeated actions)
Characteristic Item Features
• Has a goal that can be accomplished by using repeat commands (~~*forever*,~~ *repeat until*, or *repeat X times*)
Variable Item Features
• Has a math application or context   - Involves whole numbers (not fractions) [characteristic feature of this variable feature] • The specific repeat command used (~~*forever*,~~ *repeat until*, and *repeat X times*) • The specific type of repetition required (~~*forever*,~~ *repeat until*, and *repeat X times*) • The method by which repeating actions must be implemented (i.e., repeating language versus repeat commands) • Uses the scratch interface and/or code blocks

Notes. Alphanumeric codes designate the link to their respective LT statement(s). Forever loops appeared in the LT but were not used in lessons, thus they are marked with strikethrough text.