A Pair of ACES: An Analysis of Isomorphic Questions on an Elementary Computing Assessment

Miranda C. Parker
University of California, Irvine
Irvine, CA, USA
miranda.parker@uci.edu

Leiny Garcia
University of California, Irvine
Irvine, CA, USA
leinyg@uci.edu

Yvonne S. Kao
WestEd
San Francisco, CA, USA
ykao@wested.org

Diana Franklin
University of Chicago
Chicago, IL, USA
dmfranklin@uchicago.edu

Susan Krause
University of Chicago
Chicago, IL, USA
sgkrause@uchicago.edu

Mark Warschauer
University of California, Irvine
Irvine, CA, USA
markw@uci.edu

ABSTRACT

Background and Context. With increasing efforts to bring computing education opportunities into elementary schools, there is a growing need for assessments, with arguments for validity, to support research evaluation at these grade levels. After successfully piloting a 10-question computational thinking assessment (Assessment of Computing for Elementary Students – ACES) for 4th graders in Spring 2020, we used our analyses of item difficulty and discrimination to iterate on the assessment.

Objectives. To increase the number of potential items for ACES, we created isomorphic versions of existing questions. The nature of the changes varied from incidental changes that we did not believe would impact student performance to more radical changes that seemed likely to influence question difficulty. We sought to understand the impact of these changes on student performance.

Method. Using these isomorphic questions, we created two versions of our assessment and piloted them in Spring 2021 with 235 upper-elementary (4th grade) students. We analyzed the reliability of the assessments using Cronbach's alpha. We used Chi-squared tests to analyze questions that were identical across the two assessments to form a baseline of comparison and then ran Chi-Squared and Kruskal-Wallis H tests to analyze the differences between the isomorphic copies of the questions.

Findings. Both assessment versions demonstrated good reliability, with identical Cronbach's alphas of 0.868. We found statistically similar performance on the identical questions between our two groups of students, allowing us to compare their performance on the isomorphic questions. Students performed differently on the isomorphic questions, indicating the changes to the questions had a differential impact on student performance.

Implications. This paper builds on existing work by presenting methods for creating isomorphic questions. We provide valuable lessons learned, both on those methods and on the impact of specific types of changes on student performance.

ICER 2022, August 7–11, 2022, Lugano and Virtual Event, Switzerland

CCS CONCEPTS

• **Social and professional topics** → **Student assessment**; **K-12 education**; **Computational thinking**.

KEYWORDS

assessment, computational thinking, elementary education

ACM Reference Format:
Miranda C. Parker, Leiny Garcia, Yvonne S. Kao, Diana Franklin, Susan Krause, and Mark Warschauer. 2022. A Pair of ACES: An Analysis of Isomorphic Questions on an Elementary Computing Assessment. In *Proceedings of the 2022 ACM Conference on International Computing Education Research V.1 (ICER 2022), August 7–11, 2022, Lugano and Virtual Event, Switzerland.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3501385.3543979

1 INTRODUCTION

First introduced by Jeanette Wing in 2006, computational thinking (CT) is a persistent topic in today's increasingly technological society [45]. As CT is integrated into more domains and across the K-12 grade levels [24], questions arise over how to reliably assess CT and how to build valid assessments around this concept. There is no agreed-upon definition of CT [39], which means that for each CT definition, at each grade level, and within each new domain there are opportunities for even more CT assessments [40]. This is all further complicated by the time and effort it takes to make validated instruments of knowledge in computing (e.g., [29, 41]).

Previously, we embarked on developing a CT assessment to fit our definition (loops and sequences), grade level (4th), and domain (multilingual students). To evaluate the performance of students in our CT curriculum [38], we constructed and piloted a 10-question assessment on loops and sequences, the Assessment of Computing for Elementary Students (ACES) [30]. Although the reliability, as measured by Cronbach's alpha, of the assessment was in an acceptable range for early-stage research, we needed to improve the reliability if we wanted to use the assessment to more rigorously evaluate our curriculum. Reliability is important, as it can approximate the relationship among the individual items on the assessment, ensuring internal consistency [1]. For reference, the term 'items' is generally synonymous with questions, though we will discuss where the two terms diverge for our assessment in Section 3.2.

This paper reports on the next stage of our research into CT assessment creation. Assessment reliability can be improved overall by improving individual items on the assessment, as well as by increasing the number of items [1]. Cronbach's alpha is calculated, in part, by considering the number of items [11]. We set out to grow the number of items we could use by creating isomorphic versions of the original questions. We chose to create isomorphic versions of existing questions, rather than solely create new questions, because of the intentional, time-intensive efforts taken to construct the original questions from sequences and repetition learning trajectories [36]. We divided the new questions across two different test forms so that we could field test this larger number of items without overly burdening any individual student. Test forms had items in common (i.e., were identical on both test forms) as well as questions that were isomorphic. While we also worked to improve reliability by revising questions we had previously identified as problematic, those efforts are not detailed here.

This paper details the process of creating the isomorphic questions, piloting the questions on two versions of ACES, and the analysis of 235 upper-elementary students' responses across the two forms. Our research questions were:

RQ1: *What effect do different isomorphic changes have on the reliability of two forms of an assessment?*

RQ2: *What effect do different isomorphic changes have on student performance among comparable student populations?*

2 RELATED WORK

2.1 Elementary Computational Thinking Assessments

Few assessments measure computing and CT abilities among elementary students. We previously used the CSEdResearch.org database to find and review evaluation instruments that assess computing content knowledge of young children: Project Quantum [34] and Project TREES [9]. Since our original publication on this assessment [30], three additional measures can be found using the CSEdResearch database to search for elementary CT assessments. One measure is the Computational Thinking Assessment (CTA), which measures algorithms, loops, and debugging among third-grade students [42], as described by Brennan and Resnick's CT framework, ISTE, and CSTA standards [7, 15]. The CTA consists of 10 questions with five algorithmic tasks, three debugging challenges, one yes/no scenario, and one open-ended question. Though the assessment resulted in low internal reliability, they found that a one-step process as opposed to a two-step, multi-task approach was less confusing for the young age group [42]. Another measure is TechCheck, which was designed to measure CT domains (e.g. algorithms, modularity, control structures, representation, hardware/software, debugging) without requiring prior knowledge of computer programming [4, 35]. TechCheck consists of 15 multiple-choice questions composed of various "unplugged" or technology-free tasks such as puzzles, mazes, and sequencing. TechCheck resulted in acceptable criterion validity and reliability for first- and second-grade students [35]. The final assessment in the database is the Coding Stages Assessment (CSA), which measures the development of coding literacy using a five-stage learning trajectory for four- to seven-year-olds [3, 14]. As students progress through the stages, they demonstrate greater mastery of both understanding and writing code. The CSA uses 25 open-ended questions in ScratchJr to determine the coding stage of learners. The CSA questions are verbal and task-based questions, which differs from the forced-choice approach of the CTA and TechCheck.

In addition to the instruments in the CSEdResearch.org database, there are assessments found in the literature that are scoped for elementary students learning computing concepts. The Beginners Computational Thinking Test (BCTt) [46], designed for primary school students, is an adapted version of the Computation Thinking Test (CTt) [37], which was created for middle school students. Developed for upper elementary students, the CT Practices assessment [2] focuses on CT practices, rather than concepts, outlined by Brennan and Resnick [7]. There are also assessments that have been used to measure 3rd and 4th graders' understandings of sequences, conditionals, repetition, and decomposition in the context of an integrated math-CT curriculum [25].

While all these assessments address the need for evaluation tools at the primary level, there is not yet an assessment that meets our conceptual needs at the fourth-grade level.

2.2 Isomorphic Questions

Isomorphic questions are created such that certain elements are the same or similar [5]. Assessment developers often create isomorphic questions when intending to create two psychometrically equivalent test forms. Isomorphic questions have been used in computing education research to evaluate different aspects of peer instruction in computer science courses, such as investigating the effect of peer discussion on learning [33] or exploring long-term retention of learning [48]. The creation of similar, but still different, questions has also been used in the literature to mitigate cheating on CS assessments [8, 16]. Isomorphic questions were pivotal in the replication of the Foundational CS1 (FCS1) assessment to create the Second CS1 (SCS1) assessment [29]. In all of these works, the isomorphic questions kept the concept being assessed, with the SCS1 replication work also maintaining the style of the question. However, other elements, such as the text of the prompt and answer choices, are subject to changes. In computing education research, there are several different ways to refer to items that are changed but still structurally similar. Some papers only refer to this change as being 'isomorphic' [29, 33, 48], whereas other papers refer to these as 'surface feature changes' [16] or 'question variants' [8]. Due to this branching in the language around question changes, related work can be difficult to bring together on this topic. We choose to use the term 'isomorphic' as it is more commonly used in the education literature.

Using isomorphic questions to create parallel test forms was first recommended and tested by Clause et al. [10]. Lievens and Sackett [23] extended this work and defined different ways of creating isomorphic questions by applying the radicals-incidentals approach to item generation. *Radicals*, also called *controlling factors*, are item features that affect item difficulty. For a math problem, radicals might include the specific concepts being tested, the magnitude of the numbers used, and whether the problem is contextualized (e.g., a story problem) or decontextualized (e.g., solving an equation). *Incidentals*, also called *noncontrolling* or *nuisance factors*, are

item features that do not affect item difficulty. For a math problem, incidentals might include the names of students used in a story problem or other superficial elements within the story context. Using Lievins and Sackett's terminology, a pair of items is *item isomorphic* if you assume that the item's underlying constructs and the item context (i.e., the item stem and answer options) are both radicals, and only minor changes to the problem presentation are made. Another approach would be to assume that only the underlying constructs are radicals and that changes to the item stem and answer options are incidental. Lievins and Sackett referred to this as *incident isomorphic* because they were studying a specific type of item, the Situational Judgment Test, that contains a "critical incident" that respondents react to [23]. In this paper, we will refer to these types of isomorphisms as *construct isomorphic*.

Although item writers may have a priori ideas about what item features constitute radicals vs. incidentals, research suggests that it can be difficult to predict what changes will impact item difficulty. For example, Fowler et al. found that permutations to the surface feature of a question (such as changing function or variable names) did not result in significant differences in the difficulty of the question [16]. Similarly, Weston et al. found that changing the species of plant used in a biology question on photosynthesis does not change the content of the student responses [43]. These are both examples of *item isomorphisms*.

However, seemingly superficial changes can cause differences in student performance. Even if isomorphic questions result in similar scores, different amounts of time or effort may be required to answer each question [8]. Hayes and Simon investigated 13 different isomorphic versions of a Tower of Hanoi problem and found that isomorphic versions that change the context (e.g. from pegs and disks to flagpoles and acrobats) took three times as long to solve as the standard Tower of Hanoi question [19]. These differences in completion time vary even more with other changes, where some isomorphic questions took 16 times as long to complete as the original problem [19]. The effects of small changes to create isomorphic questions can also be seen in the SCS1 replication work, where some items change in their difficulty level between the original FCS1 and the replicated SCS1 assessment [29].

Although it is clear that changes in questions can create changes in performance, the underlying mechanism that causes these shifts is not always clear. Seemingly superficial changes in an item's context can cause students to recruit different knowledge and cognitive processes when solving a problem. Kotovsky et al. investigated the Tower of Hanoi problems and developed several hypotheses for the mechanisms behind the varying performance, including spatial memory load, real-world knowledge, and the use of external representations [21]. The context of a question has also been shown to have a significant impact on student performance, such as when students don't recognize that two problems, with different contexts, are fundamentally the same problem [17]. Similarly, previous studies in math education have shown that students perform differently on problems with the same underlying mathematics if one problem is contextualized and the other is decontextualized [20] or if the manipulatives used in the problem are unrealistic [27]. Kyllonen [22] wrote:

> How is it that we know which factors are radical and which incidental? One approach, and it seems the most common one, is to designate certain item factors incidental a priori.... The problem is that these a priori incidentals could turn out to be important determinants of item difficulty.... This suggests that the radical-incidental distinction could at least be partially empirically determined. (p. 261)

Thus, it is critical to analyze isomorphic questions to understand just how similar or different they are, as making seemingly minor changes can still have significant effects on student performance. In the remainder of this paper, we describe how we constructed isomorphic items, how student performance did or did not differ across pairs of isomorphic items, and what these differences might mean for constructing assessments of computational thinking.

Table 1: The questions and their response type on each form of the assessment. Questions unique to each assessment are not included in this analysis and are left blank in this table.

#	Form A	Response	Form B	Response
q3			ID1	Multi-select
q4			ID2	Multiple-choice
q5	ID1	Multi-select		
q6	ID2	Multiple-choice	ID3	Multi-select
q7	ID3	Multi-select	ID4	Multiple-choice
q8	ID4	Multiple-choice	ID5	Multiple-choice
q9	ID5	Multiple-choice		
q10	IS1	Multiple-choice	IS1	Multiple-choice
q12	IS2	Multiple-choice	IS2	Multiple-choice
q13	IS3	Multiple-choice	IS3	Multi-select
q14	IS4	Ordering	IS4	Ordering
q15	IS5	Multi-select	IS5	Multi-select
q16	IS6	Multi-select	IS6	Multi-select

3 METHODS

In this section, we first describe how we created the two isomorphic versions of the assessment. Then, we detail the study context and population. We refer to questions by "ID" or "IS" followed by a number, where "ID" stands for identical questions and "IS" stands for isomorphic questions. A full summary of the questions can be found in Table 1.

3.1 Building Isomorphic Questions

On the original assessment, each question was written to measure sequences, loops, or a combination of sequences and loops. The questions included two different prompt types: code blocks or story-style. Code block questions used blocks created via the Scratch Blocks[1] interface to make the blocks easier to interpret for 4th-grade students (e.g. using "turn right" instead of "rotate 90 degrees," such as shown in Figure 1). Story-style questions were presented in a story context, without the presence of any code in the question or answer choices (such as in Figure 6). These questions are styled

[1] https://scratchblocks.github.io/

after the Bebras tasks [12, 13] but are not a part of the official set of Bebras challenge questions.

When creating our isomorphic questions, we followed the steps taken by Parker et al. and maintained the construct being tested and how it is tested [29]. We first identified six questions that seemed easy to create isomorphisms for, which we defined as being able to change the context with few other changes. All of these questions involved navigating an animal (e.g. ladybug) to an object (e.g. leaf) on a two-dimensional (2-D) grid. Four of these were questions that used code blocks, whereas the other two questions were story-style questions. We outline the six questions selected for isomorphic adaptation and the changes made to create the isomorphic versions in Table 2, and describe each in more detail below.

We used combinations of the following changes to create the isomorphic versions of each of these six items:

Changes to the item stem

- *Grid dimensions/orientation.* We changed the size of the grid along at least one dimension. In the case of IS1, we kept the grid the same size but changed the orientation of the grid from landscape to portrait.
- *Path shape/length.* We changed the shape of the path the animal took on the grid by changing the length of one or more path segments.
- *Path direction.* We kept the path shape the same (i.e., the same number and length of path segments), but the animal traveled the path in a different direction.

Changes to the answer choices

- *Order.* We kept the answer choices the same but presented them in an alternate order.
- *Content.* We provided different answer choices for each of the isomorphs. In the case of IS4, this change was very minor, simply changing one word on one block to read "left" instead of "right" to accommodate the change in path direction.
- *Response style.* We changed the response style of one question, from a multi-select question that had multiple right answers to a multiple-choice question with a single right answer. In doing this, we also changed the number of options students had to consider.

In all cases, we changed the environment illustrations for the item, for example by taking an item that featured a sea turtle on a beach moving towards the ocean and changing it to a bunny on grass moving towards a carrot (see Figure 1). Whether these changes to the item stem and answer choices constitute *radical* or *incidental* changes (i.e., did we create *item isomorphs* or *construct isomorphs*?) will be answered empirically in Section 5.

We detail each isomorphic question below, with accompanying figures. While the images presented in this paper are static, the images on the actual assessment are animations, depicting the steps that students should mimic with the answer choice(s) they select.

3.1.1 IS1. IS1, as seen in Figure 1, is a multiple-choice question on sequences and uses code blocks. On Form A, this question involved a turtle heading across the sand to seaweed in the ocean, while Form B depicted a bunny on grass moving towards a carrot. We used this question to explore whether the orientation of the grid would impact student performance on the question. As such, Form A had a landscape grid (i.e., wider than it is tall) which meant the turtle moved from left to right more than it did vertically. Form B had a portrait grid (i.e., taller than it is wide) which meant the bunny moved more from the bottom of the grid to the top than it did horizontally. As a result of the change in grid orientation, the starting and ending positions of the sprite also changed.

3.1.2 IS2. IS2, as seen in Figure 2, is a multiple-choice question on loops and sequences and uses code blocks. The accompanying animation for this question shows the sprite moving the steps shown by the repeat block, but pausing to allow students to determine how to move towards to goal object and, consequently, what code blocks are required to make that movement. The grid orientation was the same, but the size of the grid differed between the two versions. The animals in each question also started at different positions, requiring the turtle to turn left but the bunny to turn right. As such, the steps to the goal object, and thus the correct answer to the question, differed between the two versions. We made these changes to explore if one direction might be easier than the other or if the size of a grid affected student performance on the question.

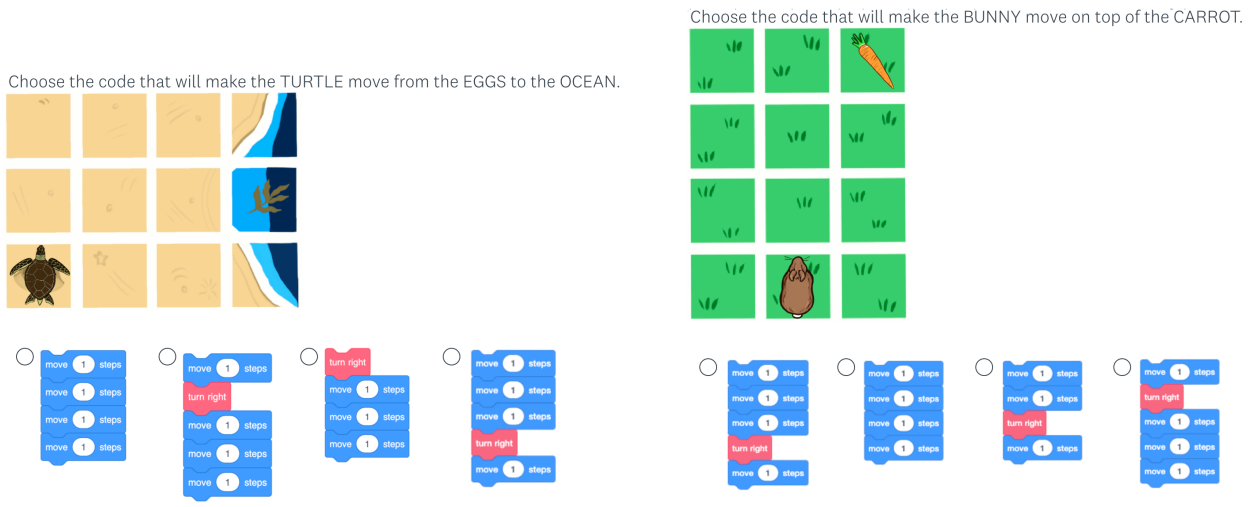
3.1.3 IS3. IS3 is a code-based loops and sequences question that can be seen in Figure 3. This question had the most significant changes between the two versions. While both questions asked the students to select the code that would make the animal move towards the object, and both questions had identical grid sizes and orientations, the questions changed how far away the object was from the animal. While just this change would have been enough to explore in analyses, we also changed the answer choices. Form A offered three choices in a multiple-choice format while Form B offered four choices in a multi-select format. One of the original intentions of the answer choices was to determine if students could identify that having a block of code before a repeat block is the same (in this instance) as having that block of code after the repeat block. However, that intent was lost amongst the changes to Form A. Due to the changes made to this question, comparing results is challenging with our scoring methods, described in Section 3.2.

3.1.4 IS4. In IS4, as seen in Figure 4, students were asked to order the code blocks to fill in the blanks in the code shown. The environment illustrations of the problem differed, but the grid size and orientation stayed the same. We altered where the sprite started on the grid, similar to question IS2, to again explore if one direction (left or right) is more difficult for students. There the turtle had to turn right to move on top of the object, whereas the bunny had to turn left.

3.1.5 IS5. IS5 was a story-style, multi-select question centered around sequences, as seen in Figure 5. Both questions asked students to navigate an animal through a sequence of shapes and colors to reach the goal object. On Form A, this question used a bee navigating through flowers to get to her hive. The bee had to travel from one corner to another, diagonal corner of the grid to reach the hive, while the bird started in the middle of a column and had to travel to a corner to reach the nest. Meanwhile, Form B used a bird navigating through clouds to get to her nest. The grid size also differed between the two forms, with Form B having a smaller grid to navigate through. Although both questions use a multi-select response style, the correct answers are not in the same place. On

Table 2: Changes made to questions to create the isomorphic versions.

Question	Changes to item stem			Changes to answer choices		
	Grid dimensions/ orientation	Path shape/ length	Path direction	Order	Content	Response style
IS1	x		x	x		
IS2	x	x	N/A			
IS3		x				x
IS4			x		x	
IS5	x	x	N/A	N/A	x	
IS6	x	x	N/A	N/A	x	



(a) IS1 on Form A

(b) IS1 on Form B

Figure 1: IS1 across the two forms, representing changes in environment illustrations, grid orientation, and starting and ending positions. Students were asked to choose the code that would move the animal on top of the object.

Form A, the second and fourth answer choices are correct. On Form B, the first, second, and fourth answer choices are correct.

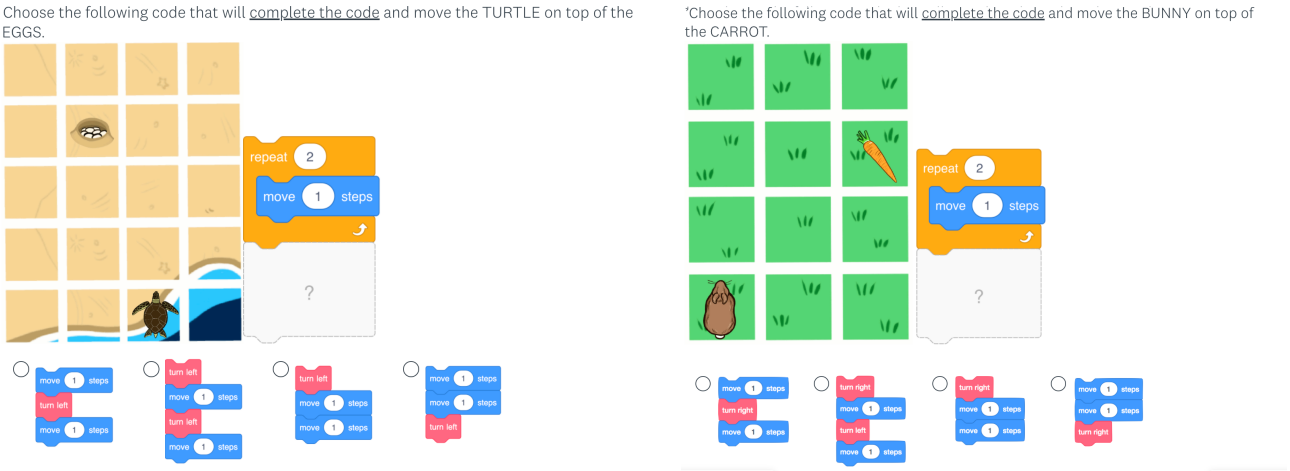
3.1.6 IS6. IS6, as seen in Figure 6, is similar to IS5, but adds in a looping notation. Again, the grid size differed between the two forms, resulting in a square grid for Form A and a landscape, rectangular grid for Form B. The starting and ending positions for the two animals also differed. The bee on Form A started at the top of the grid and need to navigate to the bottom. Meanwhile, the bird on Form B started on the left side of the grid and needed to navigate to the right. It should also, again, be noted that the answer choices on the two forms are not in the same place. The correct answers on Form A are the first, third, and fourth answer choices. On Form B, the second and fourth answer choices are correct.

3.2 Scoring Items

Since there are different response styles on the assessment, including multiple-choice and multi-select, we explored different ways of tabulating responses. After comparing scoring approaches, as described in [30], we scored the assessment based on each item, where an item may refer to the question or all available answer choices. If the question is a multiple-choice question, there is only one item since there is only one correct answer. If the question is a multi-select question, each answer choice is treated as an item. We scored items this way to ensure fairness in each correct answer choice, as well as to acknowledge that it is correct to not select an incorrect choice. We do not demerit the scores if students select an incorrect answer choice. For example, if a multi-select question has five answer choices, and only three of them are correct answers, a student who selects all possible answer choices in that question will get a total of three out of five possible points: one for each of the correct answers selected, and no points (and no negative points) for selecting the wrong answers.

3.3 Data Collection and Participant Characteristics

In order to pilot all our new isomorphic questions, we constructed two similar assessments (Form A and B) that were comprised of five identical questions, six isomorphic questions, and five unique questions. In this paper, we only discuss the questions that were identical across the two forms and the isomorphic questions. Across



(a) IS2 on Form A

(b) IS2 on Form B

Figure 2: IS2 on Form A and Form B. The environment illustrations, grid size, and direction blocks are different, but the construct, prompt, and question remain the same. Students were asked to choose the code that would move the animal on top of the object.

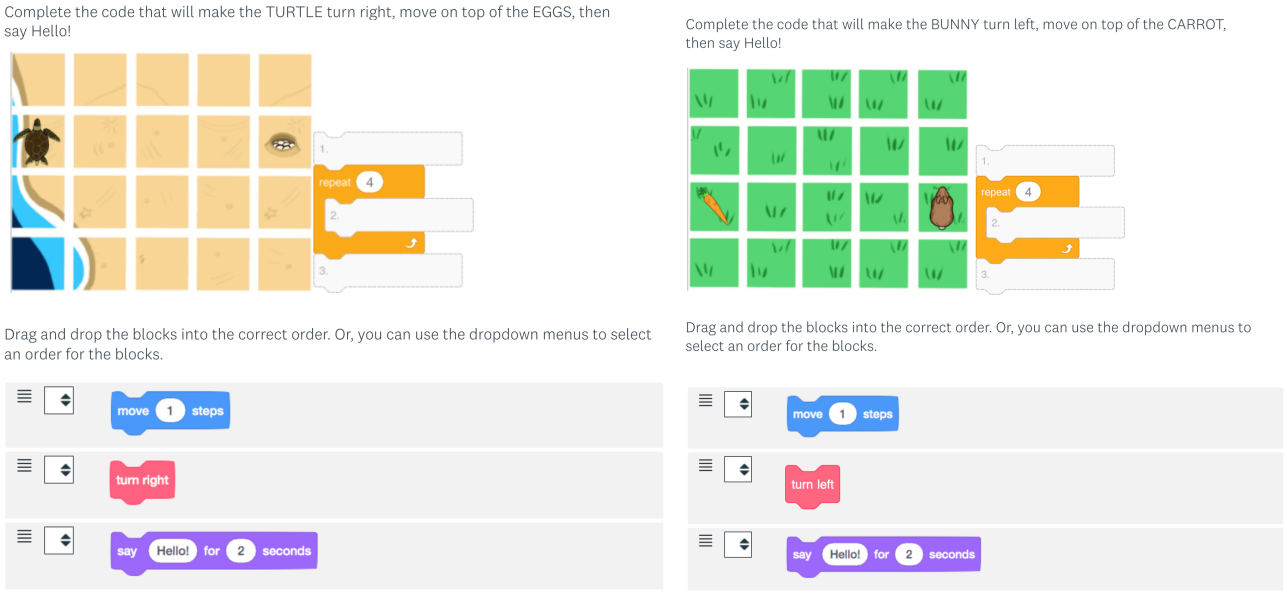
(a) IS3 on Form A

(b) IS3 on Form B

Figure 3: The two versions of IS3, representing changes in starting and ending positions of the sprite, as well as changes in the number of answer choices and response options (multiple choice and multi-select). Students were asked to choose the code that would move the animal on top of the object.

the two forms, we ensured that each assessment had a similar number of questions across each category of constructs and prompt types.

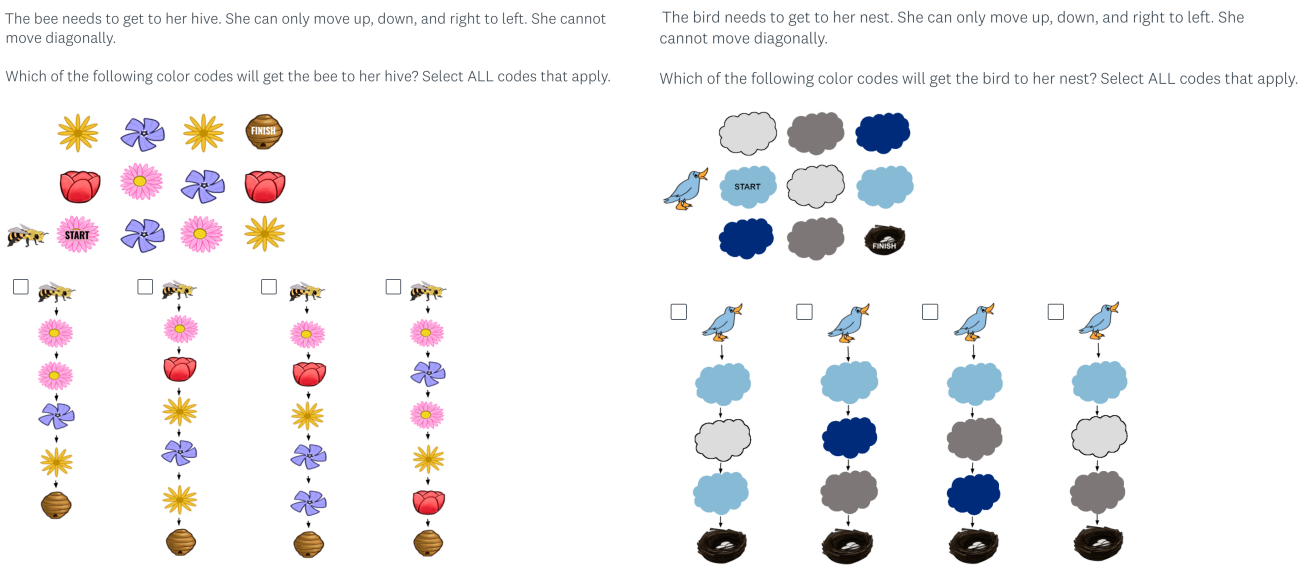
After receiving the appropriate Institutional Review Board approval for human subjects research, Forms A and B were given to schools within one school district. Five classes of fourth-grade students, representing five teachers across three schools, took Form A. Five classes of fourth-grade students, representing five teachers across three different schools, took Form B. All students, classes, teachers, and schools received the same curriculum before taking the assessment. Due to the COVID-19 pandemic, the students participated in this curriculum virtually over the course of the school year. Demographics for the students that participated can be found in Table 3.



(a) IS4 on Form A

(b) IS4 on Form B

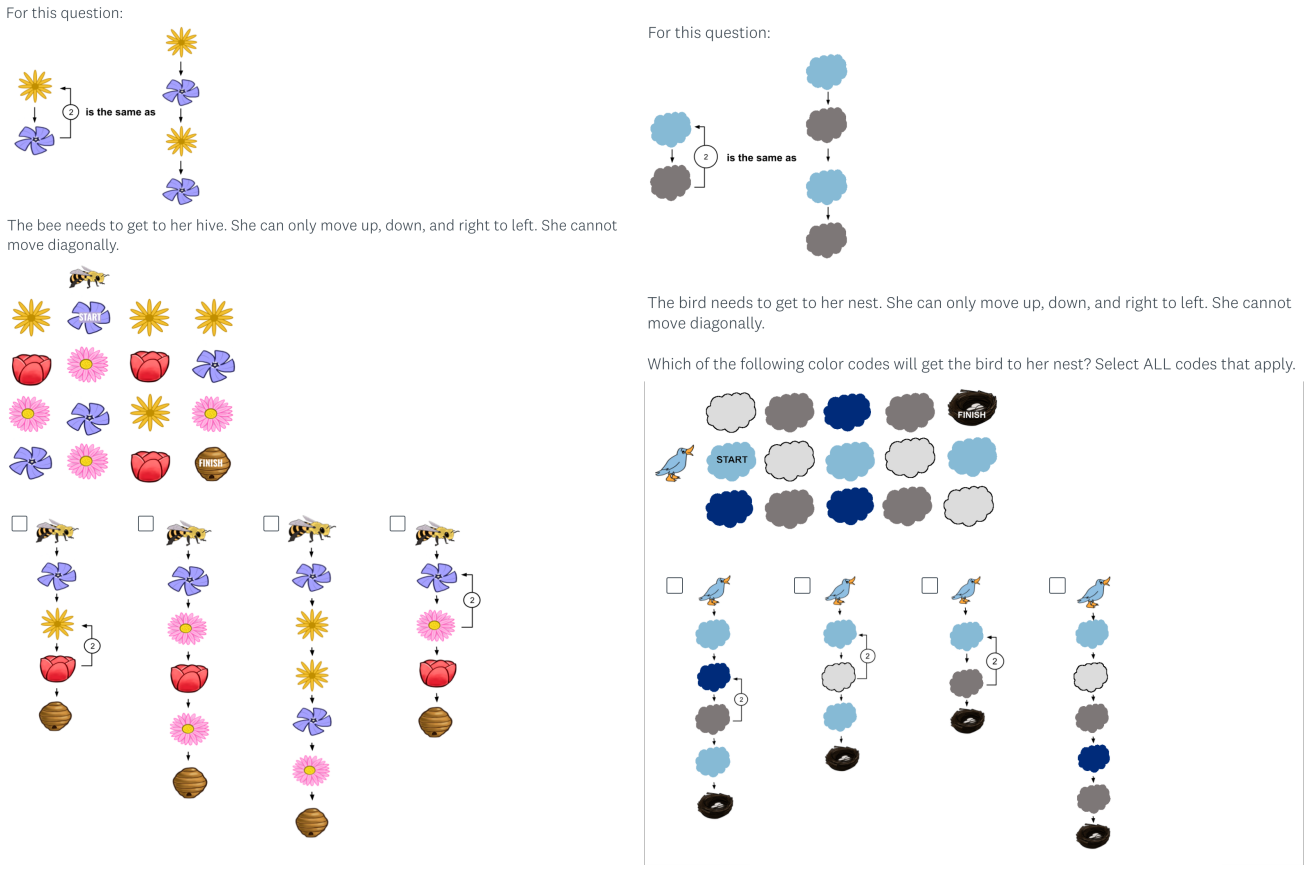
Figure 4: The two versions of IS4, representing changes in where the sprite started on the grid and thus the direction blocks used in the answer choices. Students were asked to complete the code shown next to the animation to move the animal on top of the object and then say "Hello!"



(a) IS5 on Form A

(b) IS5 on Form B

Figure 5: The two versions of IS5, representing changes in grid size. For this question, the students needed to help the animal needed to get to the object. The animal can move up, down, right, and left, but not diagonally.



(a) IS6 on Form A

(b) IS6 on Form B

Figure 6: IS6 across the two forms, representing changes in environment illustrations, grid size, and starting and ending positions. For this question, the students needed to help the animal needed to get to the object. The animal can move up, down, right, and left, but not diagonally.

Table 3: Descriptive statistics of students that completed the assessments. Some students' demographic data were unavailable. Percentages are calculated from the percent of students who took that form and had available data.

	Form A N = 113	Form B N = 122
Teachers	5	5
Schools	3	3
Data Unavailable	10	12
Girl	52 (50.5%)	61 (55.5%)
Boy	51 (49.5%)	49 (44.5%)
Hispanic or Latino	100 (97.1%)	108 (98.2%)
Designated English Learner	46 (44.7%)	36 (32.7%)
Free or Reduced Lunch	86 (83.5%)	88 (80.0%)

3.4 Data Analysis

We calculated the overall reliability for each form. For each item, we also calculated difficulty, discrimination, and assessment reliability if that item was dropped. These statistics are provided in Table 4. We analyzed the identical questions to form a baseline of comparison between our student groups. Finally, we analyzed the isomorphic questions to assess the effects of the changes made in their creation.

In Table 4, in the difficulty column, values are highlighted in green if the difficulty is greater than 0.8, indicating it is an easy question. Similarly, the value is highlighted in red if it is less than 0.2, indicating it is a difficult question. It is not inherently bad for a question to be easy or difficult, but there should be a balance of easy, moderate, and hard questions across the assessment. In the discrimination index and point-biserial correlation, values are highlighted in red if they are between -0.2 and 0.2, indicating those items do not strongly correspond with performance on the assessment overall. The Cronbach's alpha value is calculated for the assessment overall and again for each item to assess if the reliability would increase if that item were dropped. If the alpha value is higher if the item is dropped, the value is highlighted in red.

To provide us with a baseline to compare our students that took each form, we included identical questions on each form. If student scores on these identical questions are not statistically significantly

Table 4: Item difficulty, discrimination, and reliability across the three versions of the assessment. Diff. = Difficulty, DI = Discrimination index, PBC = Point-biserial correlation, Drop α = the resulting overall Cronbach's α if item were removed

Form A (α = 0.868)						Form B (α = 0.868)					
	Item	Diff	DI	PBC	Drop α		Item	Diff	DI	PBC	Drop α
ID1	ID1-A	0.57	0.37	0.32	0.866	ID1	ID1-A	0.66	0.53	0.38	0.865
	ID1-B	0.82	0.40	0.40	0.864		ID1-B	0.88	0.19	0.27	0.867
	ID1-C	0.67	0.70	0.53	0.861		ID1-C	0.71	0.38	0.31	0.866
ID2	ID2	0.43	0.27	0.16	0.870	ID2	ID2	0.52	0.53	0.35	0.865
ID3	ID3-A	0.97	0.03	0.09	0.869	ID3	ID3-A	0.98	0.03	0.10	0.869
	ID3-B	0.69	0.53	0.37	0.865		ID3-B	0.79	0.59	0.51	0.862
	ID3-C	0.68	0.40	0.34	0.865		ID3-C	0.70	0.56	0.42	0.864
	ID3-D	0.36	0.70	0.55	0.860		ID3-D	0.40	0.69	0.49	0.862
	ID3-E	0.56	0.97	0.76	0.855		ID3-E	0.61	0.75	0.54	0.861
ID4	ID4	0.58	0.50	0.36	0.865	ID4	ID4	0.57	0.44	0.34	0.866
ID5	ID5	0.42	0.90	0.69	0.856	ID5	ID5	0.50	0.75	0.54	0.861
IS1	IS1	0.84	0.30	0.34	0.865	IS1	IS1	0.57	0.63	0.43	0.864
IS2	IS2	0.42	0.60	0.42	0.864	IS2	IS2	0.65	0.63	0.46	0.863
IS3	IS3	0.61	0.47	0.38	0.865	IS3	IS3-A	0.56	0.53	0.39	0.865
							IS3-B	0.43	0.59	0.50	0.862
							IS3-C	0.81	0.31	0.28	0.867
							IS3-D	0.78	0.09	0.09	0.871
IS4	IS4-A	0.71	0.67	0.46	0.863	IS4	IS4-A	0.75	0.53	0.40	0.864
	IS4-B	0.73	0.63	0.44	0.863		IS4-B	0.75	0.63	0.45	0.863
	IS4-C	0.86	0.37	0.40	0.864		IS4-C	0.75	0.56	0.41	0.864
IS5	IS5-A	0.83	0.00	0.00	0.872	IS5	IS5-A	0.66	0.13	0.17	0.869
	IS5-B	0.64	0.63	0.49	0.862		IS5-B	0.61	0.59	0.41	0.864
	IS5-C	0.82	0.20	0.19	0.868		IS5-C	0.93	0.16	0.28	0.867
	IS5-D	0.50	0.40	0.26	0.868		IS5-D	0.45	0.59	0.39	0.864
IS6	IS6-A	0.25	0.50	0.44	0.863	IS6	IS6-A	0.80	0.19	0.16	0.869
	IS6-B	0.76	0.53	0.45	0.863		IS6-B	0.47	0.59	0.37	0.865
	IS6-C	0.60	0.53	0.37	0.865		IS6-C	0.83	0.09	0.03	0.871
	IS6-D	0.38	0.47	0.30	0.867		IS6-D	0.59	0.25	0.19	0.869

different, we can claim that the student groups are comparable, and thus their performance on the other questions is comparable. To compare student performance on these questions, we conducted Pearson's chi-squared tests of homogeneity with student dichotomous performances on items (either a student answered correctly or they did not) as our dependent variable. Our chi-square tests and associated significance values can be found in Table 5. If a chi-square value is significant, that means that student performance on that item is significantly different. The students who took Form A and the students who took Form B performed comparably on all identical questions.

After establishing a baseline of comparison among our student participants, we analyzed the student scores on the isomorphic questions. In the same fashion as the identical questions, we conducted Pearson's chi-squared tests of homogeneity with student dichotomous scores on isomorphic questions, which can be found in Table 6. In the case of IS3, which had different response styles, we compared student scores only on the similar answer choice. This meant only looking at IS3-B on Form B, which was a correct answer to IS3 and answered the question in a similar manner as the correct answer on Form A (which was the third answer choice in Figure 3a). IS3-A was also a correct answer choice on Form B but answered the question in a different manner and thus was less comparable across test forms. Additionally, for IS5 and IS6, where no answer choices corresponded across the two forms, we transformed the scores on those items to reflect an overall score on the question. Given there were four answer choices for each question, we took the average score across the items and scaled the values to between 0 and 1. Possible scores on these questions then were a set of [0, 0.25, 0.5, 0.75, 1], which meant we could not use Pearson's chi-squared tests since we no longer had dichotomous values of 0 and 1. For IS5 and IS6, we instead ran Kruskal-Wallis H tests to compare these scaled question scores, which can be seen in Table 6.

4 RESULTS

4.1 Assessment reliability

We calculated a Cronbach's alpha of 0.868 for both Form A and Form B. This indicates that each form has a reliability level that falls into the 'good' category, between 0.8 and 0.9 [28]. These values are an improvement over our previous reliability value, which was below 0.7, and thus our primary goal of growing the number of items to

Table 5: Pearson's Chi-Squared Test for Homogeneity results for the questions that were identical across the two forms of the assessment. ID stands for identical questions* indicates $p < 0.05$.

Identical Questions				
Item	Form A Mean	Form B Mean	χ^2	p
ID1-A	0.566	0.656	1.616	0.204
ID1-B	0.823	0.877	0.959	0.327
ID1-C	0.673	0.713	0.283	0.595
ID2	0.434	0.525	1.597	0.206
ID3-A	0.973	0.975	0.000	1.000
ID3-B	0.690	0.787	2.369	0.124
ID3-C	0.681	0.705	0.062	0.803
ID3-D	0.363	0.402	0.228	0.633
ID3-E	0.558	0.615	0.574	0.449
ID4	0.584	0.566	0.024	0.877
ID5	0.425	0.500	1.049	0.306

Table 6: Pearson's Chi-Squared Test for Homogeneity results for four isomorphic questions, and Kruskal-Wallis H test results for two isomorphic questions that did not have similar answer choices. The type of isomorphic question, item or incident, is also listed. * indicates $p < 0.05$.

Item	Form A Mean	Form B Mean	χ^2	p
IS4-A	0.708	0.754	0.423	0.515
IS4-B	0.726	0.754	0.121	0.728
IS4-C	0.858	0.746	3.964	0.046*
IS1	0.841	0.566	19.777	0.000*
IS2	0.416	0.648	11.740	0.001*
IS3-B	0.611	0.426	7.260	0.007*
Item	Form A Mean Rank	Form B Mean Rank	H	p
IS5	0.697	0.662	2.426	0.119
IS6	0.498	0.670	22.344	0.000*

improve the reliability was reached. There are some items that, if dropped, would increase the reliability of the assessment. These are highlighted in the "Drop α" column in Table 4, such as ID2 on Form A or ID3-A on both Forms A and B. The items that would result in the biggest changes in reliability tend to correspond with items that performed suboptimally in terms of difficulty, discrimination, and point-biserial correlation values. Each of these items will be examined to determine if they should be included as-is, edited, or removed for future iterations of this assessment.

4.2 Item difficulty

As seen in the item difficulty and discrimination analysis in Table 4, there were item-level differences across the two forms. For example, Form A has more items that were "easy" (i.e., over 80% of students answered an item correctly) but also the some of the most difficult items (e.g., IS6-A with 25% of students answering it correctly, and IS6-D with 38% of students answering it correctly). Each item with difficulty and discrimination values outside of our thresholds will continue to be examined and iterated on.

4.3 Identical items

We included identical questions to provide us with a baseline to compare student performance across the two forms. All identical items had statistically similar scores between Form A and Form B, as seen in Table 5. This is the expected result, given the students participated in the same curriculum and in the same school district and their teachers all participated in the same professional development. If scores on any of these items were significantly different, that would indicate issues in the fidelity of implementation of our curriculum across our classrooms. However, performance on these questions was similar, which demonstrates that the student populations are similar.

4.4 Isomorphic items

Due to prior work on isomorphic questions [19, 21], we anticipated that the small changes to create isomorphic questions may result in significant differences in student performance. As seen in Tables 5 and 6, performance on most of the isomorphic items (IS1, IS2, IS3, IS4-C, and an adjusted IS6) was significantly different, while only performance on IS4-A, IS4-B, and the adjusted IS5 were not significantly different. We discuss the defining features of the questions in these two groups below.

IS4 had the fewest changes – simply changing the direction of the path (and corresponding answer block) from right to left. This was the only question on each assessment that had an *ordering* response style, meaning students had to drag-and-drop answer choice blocks into the correct order. Even though the direction blocks were different, the changes between the isomorphic versions appear minimal enough to produce similar student scores for both versions of the item, though there is still a marginally significant difference in how students responded to the third blank.

Student scores were statistically significantly different on **IS1**. For IS1, we changed the grid orientation, giving Form A a landscape grid (wider than tall) and Form B a portrait grid (taller than wide), as seen in Figure 1. We also changed the path on IS1, where Form A had the turtle start in a corner of the grid and Form B had the bunny start in the center on the bottom row (see Figure 1). As a result of both of these changes, the correct answer for this multiple-choice question varied slightly across the two forms. Where on Form A the turtle needed to turn right early in the sequence of moves, on Form B the bunny needed to turn right later. Either of these factors might have played a role in student performance, given that IS1 on Form A was easier for students (84.1% of students answered it correctly) than the version of IS1 on Form B (56.6% of students answered it correctly).

IS2 had a different grid size and required different direction blocks across the two forms. On Form A, IS2 had a grid size of 5x4, where Form B had a grid size of 4x3, as seen in Figure 2. IS2 also had the animals start in different spots on the grid, which required them to move in different directions. Due to these changes, the correct answer differed between the two forms. On Form A, the correct answer involved a move block, turning, and another move block. On Form B, the correct answer involved turning, and then

two move blocks. It is possible that having a turn in the middle of a sequence of moves resulted in Form A being more difficult (41.6% of students answered it correctly) than Form B (64.8% of students answered it correctly).

As previously described, **IS3** had the most significant changes between the two versions. Thus it is no surprise that the statistical analysis confirmed that student performance on these questions was significantly different. Sixty-one percent of students selected the correct answer on Form A, whereas 43% selected the similar correct answer on Form B. However, given that Form B included another correct answer, which 56% of students selected, we can see that students may not see the move block at the end of the repeat block as a correct answer if they are presented with another correct choice (namely, the move block being before the repeat block). However, the steps needed to be taken by the animal in the question also differed. Thus, the changes we made in combination with each other made it hard to directly ascertain what affected student performance on this question.

IS5 was a sequence, story-style question. Aside from a change in the environment illustrations from a bee with flowers to a bird with clouds, the only other element that changed was the grid size (3x4 for Form A, 3x3 for Form B). However, given the path changes, the answer choices were not identical, or even similar, to each other between the two forms, which led us to compile an overall score and compare that across the forms using a Kruskal-Wallis H test (see Table 6). Overall performance on IS5 was not significantly different across the two forms, despite the changes in environment illustrations and grid size.

We analyzed **IS6** in a similar way to IS5. IS6 was a loops and sequences, story-style question that varied in the grid size and the shape of the path, as seen in Figure 6. On Form A, the animal started at the top of the grid, but on Form B the animal started on the side of the grid. As a result of these changes, the associated IS6 items were significantly more difficult for Form A students (where the correct answer choices had 25%, 38%, and 60% of students select them) than the Form B students (where correct answer choices had 47% and 59% of students select them). Generally, students could easily identify whether an answer choice was incorrect, though, with 76-83% of students correctly not selecting those choices across the two forms.

5 DISCUSSION

Every isomorphic question had a change in the environment illustrations, changing the animal and objects involved in the question. Given that this was a consistent change, there is no evidence that changing the illustrations resulted in differences in student scores. If this were the case, then every question would have been statistically significantly different. This is similar to the findings from other CS education research [16], as well as biology education research [43]. However, it is worth noting that our illustration changes were very superficial changes to the context and it may be the case that a more substantive change in context would produce differences in student performance. For example, if the isomorphic question asked students to identify the code that would replicate a specific pattern of colors, or if we decontextualized the questions and made them abstract(e.g. "Choose the code that will produce this output: x x y x x x"), it seems likely that students would perform differently. This hypothesis is in line with prior literature in this area [20, 27].

All questions had changes in the path shape (IS2, IS3, IS5, and IS6) or direction (i.e., IS1, IS4). The path change to IS4 is a simple reflection while the path change to IS1 is a rotation and a reflection. The path changes could also have affected the total distance the animal had to travel, and thus the number of move blocks needed to correctly answer the question (IS3 and IS5). The change also could have affected the placement of the turn, changing the number of move blocks needed before and/or after the turn, even if the total path length was not affected (IS2).

IS2 involved a change in the direction the sprite had to move, where the turtle turned to the left while the bunny turned to the right. However, IS4 had a similar change across the two forms but did not have a significant difference in student outcomes. This would indicate that only changing the path direction to move left vs. right does not make a significant difference. Rather, the placement of the turn (i.e. turning and then moving, as opposed to moving, turning, and then moving again) might play a role in the difference in student performance. This is similar to a previous hypothesis about spatial memory load potentially causing the outcome differences on isomorphic questions [21].

Both IS1 and IS6 had changes that ultimately affected the direction the animal had to move across the grid. In both cases, the item where the animal had to move in a left-to-right pattern was less difficult than the item where the animal had to move up and down (i.e. IS1 involved moving from the bottom of the grid to the top, while IS6 involved the opposite). Previous works have explored the relationship between the spatiality of language and spatial reasoning skills [44]. All of the students in our study are exposed to English in their classrooms, and 35% of our students are designated as English Learners, with the predominant home language for those students being Spanish. Both of these languages require students to read left to right. As such, spatial tasks that require movement from left to right may be easier for these students than tasks that move right to left [18, 47], or, in our case, up and down. These changes, and the effects on student performance, merit further investigation in future work.

As we mentioned in Section 3.1.3, we left IS3 in our analyses, despite the inherent difficulties in doing so. In effect, changing the answer choice type from multiple choice to multi-select and adding another correct answer choice made it challenging to compare the effect of each change made, especially since none of our other isomorphic questions made similar changes. However, the effects of the changes have led to some insight into when a student may identify an answer choice as correct or not in different situations. The correct answer on Form A was selected by a higher percentage of students than the similar correct answer on Form B, but Form B also had another correct answer for students to select. On Form A, students may have felt the option with the move block at the end of the repeat block was the best of the three choices and had to pick it because of the multiple-choice style. But, when presented with another option, with a move block *before* the repeat block, they are less likely to pick the option with the move block *after* the repeat block. While we will not be attempting to compare these two versions of IS3 in the future, we will continue to explore students'

conceptualizations of these two answer choices through cognitive interviews.

IS5 and IS6 are both questions that use a grid of colored shapes that an animal must move through to get to the goal object. In the instructions for these questions, students are explicitly told that the animal cannot move diagonally on the grid. However, there are answer choices presented that are only possible if the animal moves diagonally. This admittedly tricks the students into thinking that diagonal moves are possible or into forgetting that these moves are not allowed. In most cases, students easily identify these as incorrect answer choices: 80-93% of students correctly left them blank, except for IS6-C where only 60% of students left it blank. However, the inclusion of these answer choices may be testing students' ability to read and remember directions than their ability to navigate through the grid. Moving forward, we will consider the addition of incorrect answer choices that are incorrect for reasons other than this issue, to further assess the impact of the diagonal movement choices.

Taken together, the differences in student performance across the pairs of isomorphs strongly suggest that the specific spatial layout of the question is a *radical*, as discussed in Section 2.2. This would be consistent with prior work in computing education research that draws connections between spatial skills and computer science achievement [6, 26, 31, 32]. Changing the spatial layout of the questions, which often led to downstream changes to answer choices, led us to create *construct isomorphs*, which are less likely to perform similarly than *item isomorphs* [23]. The fact that student performance on IS5 did not significantly differ, despite changing the radicals of the question, is possibly due to coincidence, more than intentional design.

6 CONCLUSIONS

We set out to design and analyze isomorphic questions on an elementary CT assessment to help grow the community knowledge around building reliable assessments. After building isomorphic questions and conducting item-level analyses, we now have two versions of ACES. Each version has a better reliability measure than our original, with all reliability values in a 'good' range. With this, we have achieved our goal of growing the number of items in our question bank and improving the reliability of our assessment.

Through analyzing the identical questions on the assessments, we were able to form a baseline to compare student performance across the two forms. This allowed us to compare student performance on the isomorphic questions with the knowledge that differences are likely due to the questions and not to the student populations. Although further analysis and experimentation are needed, our current data points to illustration changes being *incidental*. Other changes, such as grid size/orientation and path shape/direction, may actually be *radical* and change an item's difficulty, even though these spatial characteristics are not central to the programming constructs being assessed.

Despite the similarly good reliability measures, we can make no claims that these assessments are equivalent or interchangeable, nor did we necessarily intend for them to be. Rather than only changing the context by changing the environment illustrations, we chose to also change other aspects of the item stem and answer choices to experiment with different factors that might affect student performance. The differences in how isomorphic questions are created can have significant impacts on how students perform on these questions. This is aligned with prior findings of the potential of seemingly small changes leading to wide variability in student performance [19, 21]. Isomorphic is not the same as identical, and thus should not be treated as such. Researchers creating assessment items should exercise caution that any changes between assessment forms may result in significant student differences. However, these changes across forms can also lend insight into what features of a question can result in the biggest changes for students. If we know that a question is easier if an animal starts in the corner of a grid, rather than in the center of the bottom row, we can use this to design more equitable assessments that truly assess computational thinking skills, rather than other confounding variables such as spatial skills or cultural background knowledge. We strongly recommend researchers conduct think-aloud interviews with students to gain greater insight about what knowledge and cognitive processes are being brought to bear when answering assessment questions, particularly if the goal is to use isomorphic items to create psychometrically equivalent test forms.

This paper contributes to the research literature terminology and guidelines for the creation of isomorphic assessment items, as well as initial analyses in an elementary CT context. Our findings may not be generalizable to all CT or computing assessments, nor all grade levels of students. More work is needed in this area to ascertain the nuances of isomorphism within different content area and contexts.

Future work on ACES includes not only analyzing student performance on questions, but also the time and effort students spend on each question, which are known to vary on isomorphic questions even if accuracy does not [8]. We plan on doing cognitive interviews with the isomorphic questions to better understand the validity of our questions and students' techniques in answering them. Eventually, we plan on conducting a large-scale psychometric analysis of the assessments to provide appropriate evidence of validity and reliability to the community, allowing for more research teams to use this assessment for their study needs.

ACKNOWLEDGMENTS

Thank you to our teachers and students who participate in this project. This work is supported in part by the National Science Foundation through Grants #1923136 and #1660871 and by the United States Department of Education through Grant #U411C190092. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the United States Department of Education.

REFERENCES

[1] American Educational Research Association et al. 2018. *Standards for educational and psychological testing*. American Educational Research Association.

[2] Satabdi Basu, Daisy Rutstein, Yuning Xu, and Linda Shear. 2020. A principled approach to designing a computational thinking practices assessment for early grades. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. 912–918.

[3] Marina Umaschi Bers. 2019. Coding as another language: a pedagogical approach for teaching computer science in early childhood. *Journal of Computers in Education* 6, 4 (2019), 499–528.

[4] Marina Umaschi Bers. 2020. *Coding as a playground: Programming and computational thinking in the early childhood classroom.* Routledge.

[5] Paul D Bliese, David Chan, and Robert E Ployhart. 2007. Multilevel methods: Future directions in measurement, longitudinal analyses, and nonnormal outcomes. *Organizational Research Methods* 10, 4 (2007), 551–563.

[6] Ryan Bockmon, Stephen Cooper, William Koperski, Jonathan Gratch, Sheryl Sorby, and Mohsen Dorodchi. 2020. A cs1 spatial skills intervention and the impact on introductory programming abilities. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. 766–772.

[7] Karen Brennan and Mitchel Resnick. 2012. New frameworks for studying and assessing the development of computational thinking. In *Proceedings of the 2012 annual meeting of the American educational research association, Vancouver, Canada*, Vol. 1. 25.

[8] Liia Butler, Geoffrey Challen, and Tao Xie. 2020. Data-Driven Investigation into Variants of Code Writing Questions. In *2020 IEEE 32nd Conference on Software Engineering Education and Training (CSEE&T)*. IEEE, 1–10.

[9] Guanhua Chen, J. Shen, Lauren Barth-Cohen, Shiyan Jiang, X. Huang, and M. Eltoukhy. 2017. Assessing elementary students' computational thinking in everyday reasoning and robotics programming. *Comput. Educ.* 109 (2017), 162–175.

[10] Catherine S Clause, Morell E Mullins, Marguerite T Nee, Elaine Pulakos, and Neal Schmitt. 1998. Parallel test form development: A procedure for alternate predictors and an example. *Personnel Psychology* 51, 1 (1998), 193–208.

[11] Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika* 16, 3 (1951), 297–334.

[12] Valentina Dagienė and Sue Sentance. 2016. It's computational thinking! Bebras tasks in the curriculum. In *International conference on informatics in schools: Situation, evolution, and perspectives*. Springer, 28–39.

[13] Valentina Dagiene and Gabriele Stupuriene. 2016. Bebras–A Sustainable Community Building Model for the Concept Based Learning of Informatics and Computational Thinking. *Informatics in education* 15, 1 (2016), 25–44.

[14] Laura E de Ruiter and Marina U Bers. 2021. The Coding Stages Assessment: development and validation of an instrument for assessing young children's proficiency in the ScratchJr programming language. *Computer Science Education* (2021), 1–30.

[15] International Society for Technology in Education and Computer Science Teachers Association. [n.d.]. https://id.iste.org/docs/ct-documents/computational-thinking-operational-definition-flyer.pdf

[16] Max Fowler and Craig Zilles. 2021. Superficial Code-guise: Investigating the Impact of Surface Feature Changes on Students' Programming Question Scores. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. 3–9.

[17] Mary L Gick and Keith J Holyoak. 1980. Analogical problem solving. *Cognitive psychology* 12, 3 (1980), 306–355.

[18] Alessandro Guida, Ahmed M Megreya, Magali Lavielle-Guida, Yvonnick Noël, Fabien Mathy, Jean-Philippe van Dijck, and Elger Abrahamse. 2018. Spatialization in working memory is related to literacy and reading direction: Culture "literarily" directs our thoughts. *Cognition* 175 (2018), 96–100.

[19] John R Hayes and Herbert A Simon. 1977. Psychological differences among problem isomorphs. *Cognitive theory* 2 (1977), 21–41.

[20] Kenneth R Koedinger, Martha W Alibali, and Mitchell J Nathan. 2008. Trade-offs between grounded and abstract representations: Evidence from algebra problem solving. *Cognitive Science* 32, 2 (2008), 366–397.

[21] Kenneth Kotovsky, John R Hayes, and Herbert A Simon. 1985. Why are some problems hard? Evidence from Tower of Hanoi. *Cognitive psychology* 17, 2 (1985), 248–294.

[22] Patrick C Kyllonen. 2002. Item generation for repeated testing of human performance. *Item generation for test development* (2002), 251–276.

[23] Filip Lievens and Paul R Sackett. 2007. Situational judgment tests in high-stakes settings: Issues and strategies with generating alternate forms. *Journal of Applied Psychology* 92, 4 (2007), 1043.

[24] James Lockwood and Aidan Mooney. 2017. Computational thinking in education: Where does it fit? A systematic literary review. *arXiv preprint arXiv:1703.07659* (2017).

[25] Feiya Luo, Maya Israel, and Brian Gane. 2022. Elementary Computational Thinking Instruction and Assessment: A Learning Trajectory Perspective. *ACM Transactions on Computing Education (TOCE)* 22, 2 (2022), 1–26.

[26] Lauren E Margulieux. 2020. Spatial encoding strategy theory: The relationship between spatial skill and stem achievement. *ACM Inroads* 11, 1 (2020), 65–75.

[27] Nicole M McNeil, David H Uttal, Linda Jarvin, and Robert J Sternberg. 2009. Should you show me the money? Concrete objects both hurt and help performance on mathematics problems. *Learning and instruction* 19, 2 (2009), 171–184.

[28] Jum Nunnally and Ira H Bernstein. 1994. *Psychometric Theory.* The McGraw-Hill Companies.

[29] Miranda C Parker, Mark Guzdial, and Shelly Engleman. 2016. Replication, validation, and use of a language independent CS1 knowledge assessment. In *Proceedings of the 2016 ACM conference on international computing education research*. 93–101.

[30] Miranda C Parker, Yvonne S Kao, Dana Saito-Stehberger, Diana Franklin, Susan Krause, Debra Richardson, and Mark Warschauer. 2021. Development and Preliminary Validation of the Assessment of Computing for Elementary Students (ACES). In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. 10–16.

[31] Miranda C Parker, Amber Solomon, Brianna Pritchett, David A Illingworth, Lauren E Marguilieux, and Mark Guzdial. 2018. Socioeconomic status and computer science achievement: Spatial ability as a mediating variable in a novel model of understanding. In *Proceedings of the 2018 ACM Conference on International Computing Education Research*. 97–105.

[32] Jack Parkinson and Quintin Cutts. 2020. The effect of a spatial skills training course in introductory computing. In *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education*. 439–445.

[33] Leo Porter, Cynthia Bailey Lee, Beth Simon, and Daniel Zingaro. 2011. Peer instruction: Do students really learn from peer discussion in computing?. In *Proceedings of the seventh international workshop on Computing education research*. 45–52.

[34] Project Quantum. [n.d.]. *Project Quantum – A Collection of Computing Quizzes*. https://diagnosticquestions.com/quantum

[35] Emily Relkin, Laura de Ruiter, and Marina Umaschi Bers. 2020. TechCheck: Development and validation of an unplugged assessment of computational thinking in early childhood education. *Journal of Science Education and Technology* 29 (2020), 482–498.

[36] Kathryn M. Rich, Carla Strickland, T. Andrew Binkowski, Cheryl Moran, and Diana Franklin. 2018. K–8 learning trajectories derived from research literature: Sequence, repetition, conditionals. *ACM Inroads* 9, 1 (mar 2018), 46–55. https://doi.org/10.1145/3105726.3106166

[37] Marcos Román-González, Juan-Carlos Pérez-González, and Carmen Jiménez-Fernández. 2017. Which cognitive abilities underlie computational thinking? Criterion validity of the Computational Thinking Test. *Computers in human behavior* 72 (2017), 678–691.

[38] Dana Saito-Stehberger, Leiny Garcia, and Mark Warschauer. 2021. Modifying Curriculum for Novice Computational Thinking Elementary Teachers and English Language Learners. In *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1*. 136–142.

[39] Valerie J Shute, Chen Sun, and Jodi Asbell-Clarke. 2017. Demystifying computational thinking. *Educational Research Review* 22 (2017), 142–158.

[40] Xiaodan Tang, Yue Yin, Qiao Lin, Roxana Hadad, and Xiaoming Zhai. 2020. Assessing computational thinking: A systematic review of empirical studies. *Computers & Education* 148 (2020), 103798.

[41] Cynthia Taylor, Michael Clancy, Kevin C Webb, Daniel Zingaro, Cynthia Lee, and Leo Porter. 2020. The practical details of building a CS concept inventory. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. 372–378.

[42] Yune Tran. 2019. Computational thinking equity in elementary classrooms: What third-grade students know and can do. *Journal of Educational Computing Research* 57, 1 (2019), 3–31.

[43] Michele Weston, Kevin C Haudek, Luanna Prevost, Mark Urban-Lurain, and John Merrill. 2015. Examining the impact of question surface features on students' answers to constructed-response questions on photosynthesis. *CBE—Life Sciences Education* 14, 2 (2015), ar19.

[44] Benjamin Lee Whorf. 2012. *Language, thought, and reality: Selected writings of Benjamin Lee Whorf.* MIT press.

[45] Jeannette M Wing. 2006. Computational thinking. *Commun. ACM* 49, 3 (2006), 33–35.

[46] María Zapata-Cáceres, Estefanía Martín-Barroso, and Marcos Román-González. 2020. Computational thinking test for beginners: Design and content validation. In *2020 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 1905–1914.

[47] Samar Zebian. 2005. Linkages between number concepts, spatial thinking, and directionality of writing: The SNARC effect and the reverse SNARC effect in English and Arabic monoliterates, biliterates, and illiterate Arabic speakers. *Journal of Cognition and Culture* 5, 1-2 (2005), 165–190.

[48] Daniel Zingaro and Leo Porter. 2015. Tracking student learning from class to exam using isomorphic questions. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*. 356–361.