Journal of Science Education and Technology
https://doi.org/10.1007/s10956-019-09805-8

Missing in Measurement: Why Identifying Learning in Integrated Domains Is So Hard

Whitney Wall Bortz[1] · Aakash Gautam[1] · Deborah Tatar[1] · Kemper Lipscomb[2]

© Springer Nature B.V. 2019

Abstract

Integrating computational thinking (CT) and science education is complex, and assessing the resulting learning gains even more so. Arguments that assessment should match the learning (Biggs, *Assessment & Evaluation in Higher Education*, *21*(1), 5–16. 1996; Airasian and Miranda, *Theory into Practice*, *41*(4), 249–254. 2002; Hickey and Zuiker, *Journal of the Learning Sciences*, *21*(4), 522–582. 2012; Pellegrino, *Journal of Research in Science Teaching*, *49*(6), 831–841. 2012; Wiggins, *Practical Assessment, Research and Evaluation,* 2(2). 1990) lead to a performance-oriented approach to assessment, using tasks that mirror the integrated instruction. This approach reaps benefits but also poses challenges. Integrated CT is a new approach to learning. Movement is being made toward understanding what it means to operate successfully in this context, but consensus is neither general nor time tested (Kaput and Schorr, *Research on technology and the teaching and learning of mathematics: Case and perspectives* (Vol. 2, pp. 211–253) 2008). Movement is also being made toward developing methods for assessing CT. Despite the benefits of matching assessment with pedagogy, there may be intrinsic losses. One problem is that interactions between the two domains may invalidate the results, either because the gains in one may be easier to measure at certain times than the gains in the other, or because interactions between the two domains may cause measurement interference. Our examination draws upon both theoretical basis and also existing practices, particularly from our own work integrating CT and secondary science. We present a mixed-methods analysis of student assessment results and consider potential issues with moving too quickly toward relying on a rubric-based approach to evaluating this student learning. Centrally, we emphasize the importance of assessment approaches that reflect one of the most important affordances of computational environments, that is, the expression of multiple ways of knowing and doing (Turkle and Papert, *Journal of Mathematical Behavior*, *11*(1), 3–33. 1992).

Keywords Computational thinking · Assessment · Learning · Science education

Introduction

This article focuses on one issue of general importance embedded in a specific, larger project that explored the integration of *computational thinking* (CT) into 7th and 8th grade earth science curriculum and classrooms: how integrating CT into classroom-based core curricular instruction should influence assessment. Best practices emphasize parallels between assessments and learning experiences, suggesting that an integrated approach to instruction should be complemented by an integrated approach to assessment. At the same time, teachers and administrators as well as researchers need to weigh the benefits of integrated instruction against existing practices and priorities. Furthermore, we need to understand the nature of student gains. These goals suggest differentiated assessment. We investigate the idea that although *assessment* may be blended, *scoring* may be differentiated. Our experience suggests that this can be done; however, it is not without consequences. We discuss some of the challenges we faced and present some things learned from taking this perspective. In particular, the design of a rubric interacts with the ability to take these dual perspectives. Indeed, under debate during development of the rubric were the trade-offs between adequate

✉ Whitney Wall Bortz
whitney8@vt.edu

Aakash Gautam
aakashg@vt.edu

Deborah Tatar
dtatar@cs.vt.edu

Kemper Lipscomb
kemperlipscomb@utexas.edu

[1] Department of Computer Science, Virginia Polytechnic Institute and State University, 2202 Kraft Dr SW, Blacksburg, VA 24060, USA

[2] Department of STEM Education, University of Texas at Austin, Austin, USA

Published online: 09 December 2019

inter-rater reliability in scoring and sensitivity to small but important movement toward incompletely mastered skills. These trade-offs had implications for the width of the gap between two epistemological stances: one that maintained that students had learned CT if they showed evidence of mastery of particulars and one that prioritized attempts to grasp the abstractions of different representations. Thus, the question of how to assess is tied to claims about learning goals and instructional strategy.

The Larger Chem+C Project

Like many projects undertaken under the rubric of design-based research in education (Barab, and and Squire 2004; Bell et al. 2004; Hoadley 2003), the current project implemented curriculum, teacher training and assessment at the same time, with the idea of iterating toward a deeper alignment of all these necessary elements. The prospects for scaling up CT rest on its ability to handle the diversity of student knowledge. The project grappled with this by working with students with little and usually no prior knowledge of CT or computing. Additionally, we saw the teachers as gate-keepers and took seriously the ideas that the teachers we worked with and hoped to engage in the future (1) had signed up to be *science* teachers, not computer scientists; (2) were working under stressful conditions, with an extremely full and challenging curriculum; and (3) despite enthusiasm, knew little initially about CT and had, at most, 3 weeks of professional development, which our observations suggest may not be enough for mastery of an entirely new field (Wall Bortz et al. 2019). Our goal was therefore to infuse elements of CT into science instruction in a way that aligned with the instructional purposes that the teachers were already engaged with.

In 2006, Wing coined the term "computational thinking" (CT) (Wing 2006) to draw attention to the habits of mind and skills associated with problem-solving like a computer scientist. This definition struck a chord both because of its pragmatic import and because computer scientists seemed to share an intuition that there was, indeed, a cognitive disciplinary core that was well described by the term "CT." Various efforts to operationalize this definition include proposed lists of elements characteristic of CT (Brennan and Resnick 2012; Lee et al. 2011; Weintrop et al. 2015). CT can be taught directly by teaching computer science, but there are many reasons to integrate it into core curriculum: to make CT available to all students in classes all students take, to alleviate pressure on an already congested K–12 curriculum and to deepen learning in both CT and the core discipline. At the same time, integrating CT into the core curriculum draws attention gaps in deep theory of learning progressions and dependencies.

In this project, we started by developing dynamic, visual simulations that illustrate important aspects of the science the students were learning. Student engagement with such simulations could by itself be considered to constitute an element of instruction in CT. However, we did more than ask students to use the simulations. We taught both CT and science by asking students to evaluate the simulations. Then, we moved into programmatic representations of the simulations. In each of the three 2-week replacement units, we asked students to find and alter progressively more complex elements of the science models in the code.

Here, we examine the responses of 182 eighth-grade students with a particular focus on those who scored the fewest points on the items that blended science and CT. We show that, in these important cases, learning in the separate disciplines may be uncovered by a differentiated approach to scoring.

In this way, while respecting the need to have efficient assessments and scoring, we bring to assessment a kind of multiplicity that acknowledges *epistemological pluralism* (Turkle and Papert 1992). Students' varied approaches to representing their learning challenged our thinking about applying traditional methods of rubric scoring to a non-traditional performance instrument.

Perspectives

Computational Thinking and Integration with Science

Despite varied approaches to defining CT and its components (Voogt et al. 2015), arguments for its applied value across disciplines are many (Gane et al. 2018; National Research Council 2011; Settle et al. 2012; Wing 2006), as are those that students should have opportunities to access CT early in their educational experience (Grover and Pea 2018; Israel et al. 2015; Lye and Koh 2014). Rather than adding to the already congested K–12 curricula, an alternate approach is to integrate CT with disciplines already offered (Qualls and Sherrell 2010; Sengupta et al. 2013; Wilensky and Reisman 2006; Wing 2006). This approach can actually deepen learning both in the discipline and CT, as this affords opportunities for more practical applications of the two (diSessa 2000; Kaput and Schorr 2008; Papert 1980; Wilensky and Stroup 1999). Science in particular shares pedagogical connections with CT (Basu et al. 2018; Dickes and Sengupta 2013; Goldstone and Wilensky 2008; Jacobson and Wilensky 2006; Reed et al. 2005). Moreover, integrating CT into classes that *all* students take may result in more widespread impact (Grover and Pea 2018; Qualls and Sherrell 2010). Even beyond this, exposure to some CT may build a foundation for later work.

Our intervention was implemented with students possessing varied levels of experience with CT, most of whom had not encountered CT concepts at all. Therefore, we viewed this work as a *bootstrapping* opportunity—introducing students to facets of CT that would act as scaffolds for future encounters with practices defined in more in-depth frameworks, such as the taxonomy of computational thinking practices in mathematics and sciences presented by Weintrop et al. (2015). This framework in particular informed the development of our assessment and thus instructional activities, as will be described in a later section. In this way, we promoted an early stage of legitimate peripheral participation. This meant *satisficing* (Simon 1955) between science and CT learning, that is, aiming to fulfill each of the multiple goals well-enough for current purposes rather than optimizing one goal. The experience of the simulation and code lay the groundwork for such concepts as (1) learning those components of syntax (objects, properties, methods, conditionals) as they expressed scientific ideas; (2) reading code; (3) planning; and (4) the process of changing, testing and debugging code. In other words, students encountered components of CT that revolved around Weintrop et al.'s (2015) modeling and simulation practices, and a raft of elements summarized by Weintrop et al. as programming, abstraction, troubleshooting and debugging. These focal elements served as overarching CT learning goals repeatedly explored by students across multiple CT topics (see Table 1). A pedagogical innovation included importing from more advanced computer science instruction the common practice of asking students to recognize bits of the science model in code they may not have been able to fully understand. This is a particularly important technique to teach explicitly in low-socio-economic status contexts where students may fear failure.

Assessing Computational Thinking

As CT makes its way into the K–12 curriculum, methods for measuring learning gains in this new domain are imperative for drawing conclusions about how it impacts learning (Lee et al. 2011; Werner et al. 2012; Grover & Pea, 2013). Existing attempts to measure CT vary; more research is needed before conclusions can be drawn about best approaches (Yadav et al. 2014).

Given the nature of CT elements as skills that one a*pplies* rather than facts that one *knows*, many have utilized performance-style assessments (Basu et al. 2018; I. Lee et al. 2011; Settle and Perkovic 2010; Sherman and Martin 2015; Weintrop et al. 2014; Werner et al. 2012). Performance-assessments match the instructional context and often involve open-ended tasks (Marzano et al. 1993). However, scoring these tasks can be time consuming (Brennan and Resnick 2012). One approach is the design of tools to automatically score students' CT through analysis of their activity in the programming environment (Koh et al. 2010; Basu et al. 2018; Moreno-León and Robles 2015). While automation reduces the workload for the scorer, it is not clear whether the system can accurately interpret the multiple presentations of students' demonstrations of CT (Brennan and Resnick 2012). Koh et al. (2014) developed a tool to assess students' applications of CT in real time. Their system displays hierarchical CT patterns based on students' coding activities and reveals these to teachers to enable formative assessment throughout an intervention. Another approach to efficiently assess CT is the design of multiple-choice assessments (Román-González 2015). Grover (2015) argues for "systems of assessment" (see also Pellegrino 2012) and refers to a study conducted in a 6-week introductory computer science (CS) course that utilized short quizzes, open-ended assignments, and a summative assessment comprised of multiple-choice and open-ended questions requiring interaction with or interpretation of code. Similarly, Meerbaum-Salant et al. (2013) balance conclusions from one assessment of CS concepts with other qualitative measures, a balance that Yaşar (2018) argues is necessary for understanding the impact of CT in education. Blikstein and Wilensky (2009) utilize multiple forms of data in order to understand students' CT learning, including observations, student interviews, questionnaires, open-ended modeling tasks, and other student-generated artifacts.

The majority of CT assessments assess only CT, but with CT now seeping out of the CS classroom into other disciplinary areas, we face the challenge of assessing students' experiences learning two integrated domains. In fact, the challenge is not limited to the assessment of CT, as both the *Framework for K–12 Science Education* (Council, Education, Education,, and Standards 2012) and the Next Generation Science Standards (NGSS Lead States 2003) specify the integration of cross-cutting concepts and skills across science education. Pellegrino (2013) argues that given the integrated and cross-influential nature of assessment and instruction, there is a need for measurement tools that integrate these concepts with the discipline. The complexity of CT itself makes integration particularly challenging, as CT has been conceived of as comprising multiple skills, and CT knowledge and programming knowledge can also be difficult to differentiate (Koh et al. 2014).

Few reports of such integrated assessments exist. Perković et al. (2010) describe integration of CT into various general education courses at the undergraduate level for which they developed integrated assessment tasks. They acknowledge that further refinement of such tasks is needed. Weintrop et al. (2014) report on authentic assessments that mirrored the instructional context of CT and STEM. These are administered online and in the programming environment, thus allowing students to "explore and engage with the concepts in a dynamic way" (p.25). Basu et al. (2017), rather, include both science and CT concepts on their pre-/post-test, but each domain is assessed in isolation without any tasks that required integrated application of the two. More research is needed to determine how and whether to integrate this cross-cutting concept with disciplinary learning.

Scoring assessments of CT can be particularly problematic with varying understandings of the concept being evaluated. Creating rubrics to guide analysis can add reliability to the scoring process but also poses an arduous task. Sherman and Martin

Table 1 CHEM+C curricular topics by CCT

	CCT1	CCT2	CCT3
Topic	Water formation and splitting	Copper sulfate and aluminum reaction	Natural carbon cycle
Driving questions (Krajcik et al. 1998)	"What is happening that we do not normally see? What is present but is not shown by the model?"	"How do you know that a chemical reaction has occurred?"	"How does carbon move through the environment?"
Anchoring phenomenon (Schwarz et al. 2009)	Students perform physical demonstration of the electrolysis of water using a battery submerged in a tub of water with test tubes above each terminal to collect hydrogen and oxygen gas.	Teacher performs physical demonstration of the aluminum/copper sulfate reaction. An aluminum wrapped thermometer is submerged in a beaker of copper sulfate and water, resulting in the displacement of copper as a precipitate.	The simulation is used as the anchoring phenomenon with settings restricted to show only macrophenomenon activity in the carbon cycle.
Science topics	Law of conservation of mass, balanced chemical equations, properties of a chemical reaction, collision theory, reversibility, molecules vs. atoms, and purposes and types of models (NGSS: PS1.A, PS1.B, PS3.B, MS-PS1–5, MS-PS1–4, MS-PS1–2, MS-PS1–5, MS-PS1–2)	Review of CCT1 topics, properties and signs of a chemical reaction, atomic theory, precipitate, stoichiometry, chemical vs. physical change, and limiting reagents (NGSS: PS1.A, PS1.B, PS3.B, MS-PS1–5, MS-PS1–4, MS-PS1–2, MS-PS1–5, MS-PS1–2)	Review of CCTs 1 and 2 topics, carbon cycle, nature of carbon, photosynthesis, cellular respiration, energy transformation (NGSS: LS1.C, LS2.B, PS3.D, MS-LS2_1, MS-LS-1-7)
CT topics	Basics of a NetLogo model, objects, properties, breeds, debugging, and model evaluation	Review of CCT1, coordinate plane in NetLogo, patches, and conditionals	Review of CCTs 1 and 2, object positioning, object movement, variables and names, and function call and statement execution
Simulation model improvement	Students add an object (Epsom salt) to the model and assign scientifically appropriate properties	Students make the background color dependent on the amount of copper sulfate.	Students add worms or mushrooms (detritivores) that decompose plants or animals.

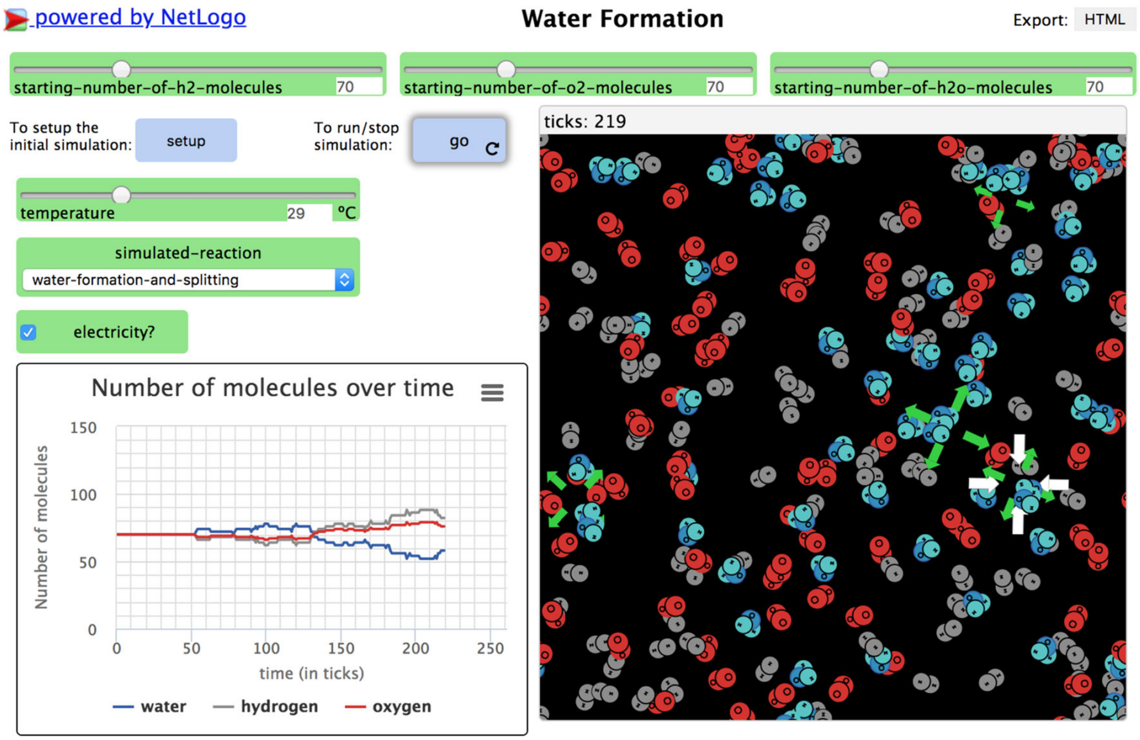
(2015) describe an iterative process of use and revision of a descriptive rubric to evaluate students' application designs for mobile CT, a subset of CT. Werner et al. (2012) share a similar approach to rubric-based evaluation of students' programming tasks in the Alice programming language/environment. Our work also includes rubric-based scoring, but we augment the research on CT assessment by describing an assessment and rubric that integrates science and CT as well as important lessons learned through a mixed-methods analysis of the results.

The CHEM+C Intervention

The research described here is part of a multi-year, multi-school study of disciplinary integration of CT—in this case, with middle school (seventh- and eighth-grade) science. The intervention consisted of 3-week-long curricular sequences, referred to as computational chemistry tasks (CCTs), designed to replace regular eighth-grade instruction of chemistry concepts. The research team designed each CCT to address chemistry concepts with which students typically struggle, such as chemical equilibrium (Sendur et al. 2010), and the nature of matter (Herrmann-Abell and DeBoer 2011; Lee et al. 1993; Stavy 1991)—concepts that also present in the Texas Essential Knowledge and Skills (TEKS) for science (Texas Education Agency 2013), and the Next Generation Science Standards (NGSS) (NGSS Lead States 2003). These science principles were modeled in NetLogo.

Curriculum and Technology

Six teachers in three different schools participated in the intervention, which in the first year was led by researchers as a pilot. In the second year, researchers observed and collected video data while teachers taught the CCTs. Our approach to integrating CT assumed that students and teachers had no prior experience with CT or its integration, yet the goal was to provide students with a CT-infused approach that prioritized deepening their understanding of the science. Each CCT followed a curricular sequence that combined traditional science strategies, such as physical experiments and argument-driven inquiry (ADI) (Sampson and Grooms 2008), with specially designed technological environments implemented in NetLogo (Wilensky 1999). Figure 1 presents the interface for the first CCT. On day 1 of each CCT, an *anchoring phenomenon* (Schwarz et al. 2009) introduced the science to be learned (Table 1). These demonstrations served to anchor the students' investigations in the computational model, which showed the underlying molecular phenomenon that gave rise to the observed macro-level phenomenon. After initial exploration of the NetLogo model using only the



Water Formation

starting-number-of-h2-molecules 70

starting-number-of-o2-molecules 70

starting-number-of-h2o-molecules 70

temperature 29 °C

simulated-reaction: water-formation-and-splitting

☑ electricity?

ticks: 219

Fig. 1 NetLogo model of water forming and splitting

simulation and not yet evaluating the code, student groups drew “another kind of model,” depicting on a white board what they thought was happening in the computer simulation, noting its key elements. Groups presented these to the class and received feedback from the teacher and their peers (a typical ADI activity). The transition from macro-level demonstration to micro-level model emphasized the nature of models as useful but imperfect (Box 1976, 1979). This contextualized a model evaluation activity in which, using a guided worksheet, students were asked to identify visible objects and their properties in the simulation and call out aspects that were either scientifically accurate, inaccurate, or missing. Class discussions of these ideas led to scaffolded code modifications intended to improve the scientific accuracy of the model. Teachers guided students’ as they utilized existing code as a model for modification. Prior to making changes to the model, the students planned changes by filling out a graphic organizer or writing pseudocode. This instruction progressed in depth in each CCT. We note that sometimes teachers stepped outside the science model to illustrate such concepts as truth values nested in conditionals. For example, some teachers started using a form of pseudo-code to organize classroom behavior, as in “if you are wearing sandals and a t-shirt, get your computer.”

Prior to implementing changes, students worked in groups of three or four to plan how they would change the code. Groups then had the opportunity to provide one another with constructive criticism on their plans (a key element of ADI). Changes to the model code were supported by the instructor or by peers who successfully completed the task. Students were asked to write reflections of their learning of science and CT during each CCT.

An example of the integration of science learning and CT in CCT1 was asking students to plan and implement the presence of the $MgSO_4$ (Epsom salt) that they had seen in the anchoring phenomenon and identified as missing from the simulation. As a first exercise, this did not involve the complex chemistry of catalysts, but it did involve debate about shape, size, distribution, and behavior of new objects. Some argued that $MgSO_4$ should be bigger than H_2O because it had more molecules; others argued that it should be smaller because it did not play a role in the parts of the process we were simulating. Later in CCT3, because students had difficulty balancing the chemical formula for the breakdown of glucose, we asked them to derive it by locating the model in the code. Students expressed surprise that when they saw the same number of elements in the inputs and outputs of the code, and teachers expressed surprise that the students were surprised. Table 1 provides additional detail on the content of topics covered in each CCT. The immensity of topics covered reflects the provision of a rich learning environment (Järvelä 1995). While each topic was not explicitly “taught,” the intervention provided students with opportunities to deepen learning across these topics, many of which were covered in prior science instruction.

Participants and Data Collection

We focus here on data from the classes of the three eighth-grade teachers in a suburban area of low to moderate socio-economic standing near Austin, TX. Informed consent was obtained from all teachers and students included in the study. Participants included the 182 students in these teachers' classes, who all consented to participate and were present for both the pre-/post-administrations of the performance assessment. At the end of the project 39.5% of students indicated that this project was their first experience with coding.

Sources of data include student-generated artifacts, such as images of group drawings interpreting the *anchoring phenomenon*, individual worksheets completed when exploring and critiquing the model, student-written reflections on their learning, video data from one class for each teacher, a pre-/post-computational attitudes test (CAT), and a pre-/post-performance assessment of computational thinking (PACT). Our focus is the pre-/post-performance assessment, but other data are described elsewhere (Gautam et al. 2017; Wall Bortz et al. 2019).

Design of the PACT

Because of the reflexive relationship between instruction and assessment (Biggs 1996), as science instruction evolves with the integration of twenty-first century skills (Pellegrino and Hilton 2013), assessments must follow. The PACT, as reported here, was under iterative construction in relationship to the curriculum. We note that the larger argument about the approach to scoring is, by and large, independent of its goodness. Nonetheless, the PACT consisted of five multi-component tasks that were thought to allow students to demonstrate both science learning and learning related to programming and/or CT. Gane et al. (2018) argues that assessments act as statements of what we want students to know and to be able to do. However, many science assessments prioritize factual knowledge over conceptual understanding (Herrmann-Abell and DeBoer 2011). We aimed to design a performance-based assessment on which students could demonstrate conceptual understanding of both CT and science. We wanted students to apply these domains in tandem so we chose to assess them with one instrument. This choice was also influenced by literature arguing for authentic assessments (Wiggins 1990) that match instruction (Airasian and Miranda 2002; Hickey and Zuiker 2012) and, as a result, reflect the integration of core disciplines, such as CT (Pellegrino 2013). "Matching to instruction," however, becomes complicated when (1) the intervention changes the underlying epistemological infrastructure of "what it means to know a subject" (Kaput and Schorr 2008) or (2) as we will demonstrate, the student must have gained in all integrated domains to be regarded as having learned at all.

Design of the PACT as administered to these students utilized an evidence-centered design (ECD) (Mislevy and Haertel 2006) process with some progress in iterative validation (Buffum et al. 2015). First, domain analysis included the design of several performance expectations (PEs; Table 2) leading to the creation of items on the written assessment through three major phases. Each PE identified a core idea in science, a scientific practice, and a CT practice we wished to target based on the NGSS (NGSS Lead States 2003) and Weintrop et al.'s (2015) taxonomy. Second, we determined what kind of student data could provide evidence that students knew the core ideas and were able to do the practices. Third, we formulated a task that would generate those kinds of data.

We constructed six PEs and presented them to participating teachers and to an advisory board composed of experts in Chemistry, CS, and STEM Education. We asked for feedback regarding to what extent the core ideas and practices we identified were worth targeting, how well the data we proposed would provide evidence about whether the expectations were met, and how well the proposed task would provide students with an opportunity to produce the data we suggested. We also piloted a draft of the assessment with 70 eighth graders who were asked to highlight any parts of the assessment that were unclear. Based on input from these parties and the need to administer the test in 40 min, we narrowed our scope to three PEs and thereby constructed the assessment items. The PEs follow: (1) "Given a section of code for an agent-based computational model of a chemical reaction, identify the properties and actions of an object"; (2) "Explain how to change an agent-based computational model of a chemical reaction to implement a similar but distinct chemical reaction"; and (3) "Given a section of code for a procedure representing a chemical reaction in an agent-based computational model, extend the code to implement a procedure for a similar but distinct chemical reaction." See Table 2 for a fuller explanation tied to the assessment item analyzed later.

The assessment included items requiring students to (1) interpret excerpts of code, and at times, what actions they would produce in a model; (2) answer questions about the science concepts encountered in the CCTs; and (3) indicate modifications they would make to an excerpt of code to implement a different chemical reaction and justify these changes. Due to limited resources of the research team for design and scoring as well as technology resources at the school, the PACT was administered as a paper-pencil task.

We secured an additional research site for the purpose of further piloting the curricular sequence and the PACT. Researchers led the three CCTs with two seventh-grade science classes and administered the CAT and the PACT before and after the units. Students' questions about and performance on the items influenced modifications to the wording of the instructions as well as changes to some of the items. Our iterative design process reflects an exploratory approach to

Table 2 Example performance expectation

<table>
<tr><td colspan="3"><i>Performance Expectation #3</i>: Given a section of code for a procedure representing a chemical reaction in an agent-based computational model, extend the code to implement a procedure for a similar but distinct chemical reaction.</td></tr>
<tr><td><i>Chemistry Core Idea</i>: Substances react chemically in characteristic ways. In a chemical process, the atoms that make up the original substances are regrouped into different molecules, and these new substances have different properties from those of the reactants. The total number of each type of atom is conserved, and thus the mass does not change.</td><td><i>Computational Thinking Practice</i>: Constructing computational models: Create new or extending existing computational models by encoding model features in a way that a computer can interpret</td><td><i>Science Practice</i>: Constructing explanations and designing solutions: Apply scientific ideas or principles to design; construct; and/or test a design of an object, tool, process, or system.</td></tr>
<tr><td colspan="3"><i>Potential Data</i>: A functional section of code that will faithfully represent the chemical reaction requested, including proper stoichiometry and identification of reactants and products</td></tr>
<tr><td colspan="3"><i>Potential Task</i>: The respondent is given a functioning section of code for an agent-based computational model of the water decomposition reaction and the balanced equation for the decomposition of glucose. The respondent is then asked to write the code to implement the glucose decomposition reaction.</td></tr>
</table>

the development of an integrated assessment—a new endeavor for both the teachers and researchers involved—with the purpose of making moves toward a more sophisticated final product.

Results

Assessment Scoring and Inter-rater Reliability

Due to time and personnel constraints, the assessments were to be scored by undergraduate researchers who were neither science educators nor computer scientists. Therefore, a clear and reliable rubric was imperative (Miller and Linn 2000). Using an iterative process of design, discussion, application and redesign of rubric levels and criteria descriptors, the research team eventually agreed upon a dichotomous rubric that broke each task down into multiple criteria and assigned either a 0 (not demonstrated) or a 1 (demonstrated) to each, resulting in more ratings. It increases reliability (Moskal 2000) across raters by reducing subjectivity in the scoring process and is useful if the raters have differing background knowledge. Once the research team was able to come to agreement on scoring using the rubric with a sample of pre- and post-assessments, inter-rater reliability was investigated with one undergraduate research assistant and one member of the research team. These two raters scored a sample of 170 assessments (31% of the total), including both pre- and post-assessments and from all participating teachers' classes. A Cohen's kappa analysis was run on each rubric item. After the first round of scoring, kappa scores for 18 of the 53 rubric items were below 0.8, with 10 items scoring below 0.6. The research team discussed these items and modified the rubric to provide more clarity to the scoring process. The two raters then scored 10 more assessments together and 20 assessments separately, and the Cohen's kappa analysis was repeated for each rubric item. The kappa scores for only four items fell below 0.8 and one item scored below 0.6. Raters met to reconcile differences on these five items and negotiated a shared process moving forward. Some items required a different approach to scoring while others benefited from modifications of the rubric to prevent disagreement. The first line of Table 3 shows considerable gains from pre- to post-test, and the temptation might be to stop analysis there.

Domain-Specific Investigation

However, when two domains are integrated, it is not as if the learning objectives no longer exist in isolation. Indeed, teachers and administrators as well as researchers want to know where students excelled and where they struggled in relation to the science and the CT separately. We also wanted to investigate whether the application of one may occur in tandem with, or even enhance or hinder, the learning of the other. We anticipated that as cultural and linguistic literacy can skew assessment results (Abedi 2002), proficiency levels in one domain of learning on an integrated assessment item may affect the representation of student learning in the final results.

To further explore students' performances, two researchers independently coded each item (Moskal 2000) on the dichotomous rubric according to the domain of learning addressed and then came to consensus through discussion. For these purposes, we treat programming and CT as a single category (Lye and Koh 2014) while acknowledging arguments for their independence (Bell et al. 2015). Because this intervention grounded the integration of CT in a particular programming environment, the assessment tasks and these analyses mirror that linkage. Therefore, each rubric item was labeled as measuring either science (S) in isolation, CT in isolation, or science and CT together (S/CT) together. An example science item asked students to write the chemical reaction equation for the formation of water from hydrogen and oxygen, while an

Table 3 Pre-/post-assessment gains by domain of learning

N = 182	Pre-test median score	Post-test median score	*P* value	Effect size
Overall *53 points possible*	4.0	18.0	< 0.001	0.825
CT *20 points possible*	3.0	12.0	< 0.001	0.768
Science *10 points possible*	0.0	3.0	< 0.001	0.736
Blended items *23 points possible*	0.0	3.0	< 0.001	0.635

example CT item asked students to identify the objects and their properties from an excerpt of NetLogo code. An example of a blended science and CT item is presented in Fig. 2.

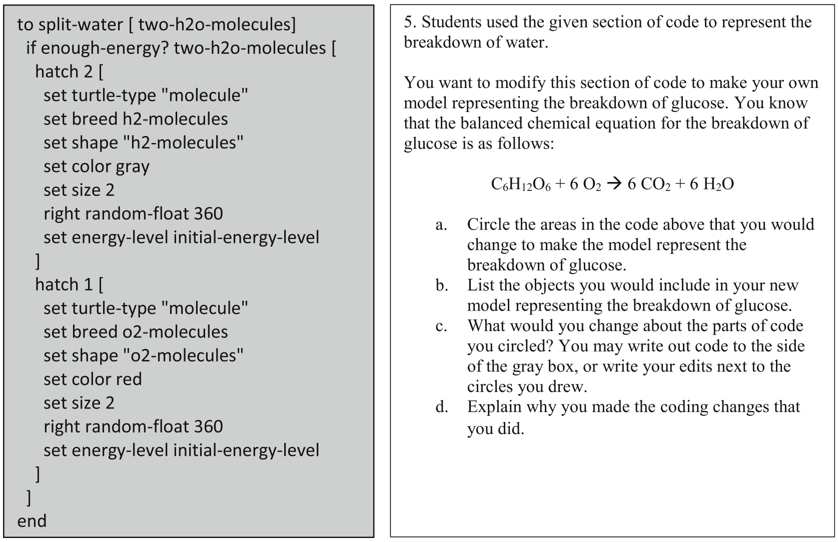
Of the 53 lines of the scoring rubric, 20 items were coded as CT, 10 as S, and 23 as S/CT. Table 4 presents examples of each. We were specifically interested in looking at incorrect responses on the items that blended CT and science to see whether evidence of learning in either science or CT could have been masked by an incorrect application of the other. We suspected that a student may have learned something; however, failure to apply learning in both domains could result in a score of "0" so that this learning would not be revealed within the quantitative analyses of scores. The following section describes both quantitative and qualitative analyses in order to highlight the evidence that was captured through a qualitative analysis but not uncovered through rubric scoring.

To measure whether the change in pre- and post-test scores was statistically significant, we conducted a Wilcoxon signed-rank test, appropriate for non-normally distributed data, to analyze differences in means between the pre- and post-test scores, thus treating each administration as a repeated measure on the same sample of students. Assessment scores are reported in Table 3. Gains were proportionally greatest on CT items and lowest on the items that blended science and CT concepts. Concern over students' poor performance on these items led to a qualitative analysis of their responses.

Qualitative Analyses

Traditional scoring of a performance assessment involves a rubric that specifies expected demonstrations of learning by students; using such a rubric can produce more reliable scores (Miller and Linn 2000). Integrating CT is a relatively new endeavor. Our aims were both to evaluate the sensitivity of our assessment and accompanying scoring method and to learn more about what students gained from the intervention.

To focus the qualitative analysis, we selected the students who had scored the lowest on the items that blended CT with science (totaling 0 to 3 out of 23 points possible). A total of 97 students (out of 182) met these criteria with 42 students scoring 0 points, 19 scoring 1, 25 of them scoring 2, and 11 achieving 3 out of the 16 possible points on the blended items. Two researchers engaged in joint qualitative analysis (MacQueen et al. 1998), looking at students' responses to blended S/CT tasks on item five (Fig. 2). Using an inductive coding process (Thomas 2006), we annotated the responses and noted areas where we observed students demonstrating



```
to split-water [ two-h2o-molecules]
 if enough-energy? two-h2o-molecules [
  hatch 2 [
   set turtle-type "molecule"
   set breed h2-molecules
   set shape "h2-molecules"
   set color gray
   set size 2
   right random-float 360
   set energy-level initial-energy-level
  ]
  hatch 1 [
   set turtle-type "molecule"
   set breed o2-molecules
   set shape "o2-molecules"
   set color red
   set size 2
   right random-float 360
   set energy-level initial-energy-level
  ]
 ]
end
```

5. Students used the given section of code to represent the breakdown of water.

You want to modify this section of code to make your own model representing the breakdown of glucose. You know that the balanced chemical equation for the breakdown of glucose is as follows:

$$C_6H_{12}O_6 + 6\ O_2 \rightarrow 6\ CO_2 + 6\ H_2O$$

a. Circle the areas in the code above that you would change to make the model represent the breakdown of glucose.
b. List the objects you would include in your new model representing the breakdown of glucose.
c. What would you change about the parts of code you circled? You may write out code to the side of the gray box, or write your edits next to the circles you drew.
d. Explain why you made the coding changes that you did.

Fig. 2 Blended S/CT question in our PACT

knowledge of either science or CT that was not captured by the numeric score. Our qualitative analysis also identified important alternative conceptions (Herrmann-Abell et al. 2016) in one domain that hindered their ability to perform in the other domain, thus leading to an incorrect response to the integrated task. Such analyses can allow the assessment to be used as a powerful formative tool (Basu et al. 2018).

We focused our analysis on responses to the blended question in the PACT (Fig. 2). The first four items on the assessment included simpler tasks, such as object and property identification in code, writing a balanced equation, and naming a scientific law. The fifth item presented both science concepts and code with which students had worked in CCTs 1 and 3, requiring a high level of science and CT integration. Presented with the code for splitting water, students were asked to follow Several Steps Toward Changing the code to represent the breakdown of glucose and to offer a justification. Performance on item five was broken down into 23-line items on the rubric. Table 4 presents one example aligned to each of the three domains.

Learning Uncovered

The qualitative analysis of students' responses on the blended S/CT tasks uncovered examples of two key aspects of student learning that the researchers agreed could not be detected through rubric scoring. First, difficulty in one aspect of the two domains (S/CT) sometimes challenged students' representations of learning. As a result, students sometimes demonstrated learning differently from that specified in the rubric, resulting in a score of zero. Second, analyses also revealed the presence of alternative conceptions that would be important for consideration by a researcher or teacher when planning subsequent lessons or interventions.

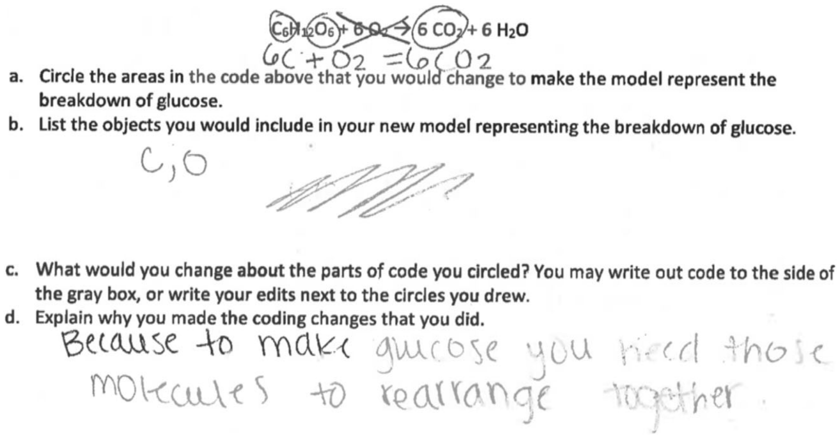
Figure 3 shows both types of undetected learning. This response did not receive any points, because S1 circled the chemical equation provided in the question rather than the code, and instead of representing the breakdown of glucose in code, the student changed the equation to what appears to be an incorrect attempt at an equation for the formation of carbon dioxide. On the pre-assessment, S1 had simply written "IDK" for the entire task. While the score did not increase, we can still learn from this student's approach on the post-test. First, the lack of vocabulary in one domain (CT) affected the demonstration of performance in the other. The term "code" was used throughout the CCTs in the programming context, but here, S1 misinterpreted "code" to refer to the chemical formula rather than the snippet of NetLogo code. Many others ($n = 42$) also interpreted the chemical formula as the "code" and attempted to change the equation in a way that made sense to them (Fig. 4 and Fig. 5). Although the word code is a term of art in programming, it specializes a more familiar, general concept. In some sense, a chemical formula is a code in which terms stand for objects, organizations of objects, and operations in the world.

While S1 was inhibited by a misinterpretation of "code," there were aspects of CT that the student *did* demonstrate. A correct response to 5b would have included the objects as indicated by the chemical equation for the breakdown of glucose (glucose, oxygen, carbon dioxide, water). However, since a misunderstanding of "code" led S1 to offer a different equation, naming those objects was no longer appropriate. Instead, S1 named the two objects ("C" and "O") present in the new unbalanced equation provided in 5a. This indicates an understanding of "objects" that would be included in a model were the student to create a computational model of this chemical formula, albeit not at the level that these models were constructed. In this case, the response to 5b depended on performance in 5a, and to that extent, the student demonstrated an understanding of the objects (atoms in this case) that would be necessary in their computational model. The student's misinterpretation of "code," that is, misunderstanding the science involved, prevented the scorer from seeing aspects of CT learning. Task 1 on the assessment evaluated CT in isolation, and when asked to identify objects in the model, the majority of students, including S1, identified the objects correctly. By contrast, in the context of an integrated task, the student's incorrect approach to the science hindered their performance on the CT, as specified in the rubric.

S1's response also revealed an alternative conception. An earlier question on the assessment asked students to state or describe the scientific law or theory that supported a balanced

Table 4 Example criteria from the rubric for scoring task 5

Task	Rubric criteria	Domain
5a–c	Did the student include accurate stoichiometry to represent the chemical equation? *6 used before the carbon dioxide molecule*	Science
5d	Did the student justify the changes necessary to represent the chemical reaction model? *Procedure name: to properly communicate the function of the procedure*	CT
5a–c	Did the student correctly form a list representing all reactants in line 1? *The list in line 1 includes the following: glucose-molecule, first-oxygen-molecule, second-oxygen-molecule, third-oxygen-molecule, fourth-oxygen-molecule, fifth-oxygen-molecule, sixth-oxygen-molecule or six-h2o-molecules*	Science/CT



$C_6H_{12}O_6 + \cancel{6\,O_2 \rightarrow} 6\,CO_2 + 6\,H_2O$

$6C + O_2 = 6CO_2$

a. Circle the areas in the code above that you would change to make the model represent the breakdown of glucose.

b. List the objects you would include in your new model representing the breakdown of glucose.

C, O

c. What would you change about the parts of code you circled? You may write out code to the side of the gray box, or write your edits next to the circles you drew.

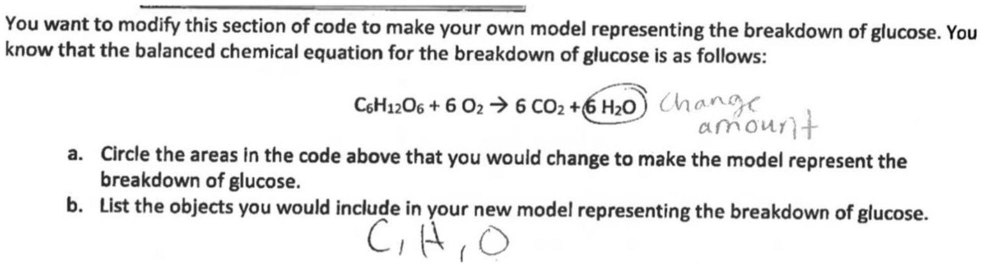
d. Explain why you made the coding changes that you did.

Because to make glucose you need those molecules to rearrange together.

Fig. 3 S1 identifies objects for an incorrect equation

equation for the formation of water. S1 received full points for the statement, “the theory is that nothing can be created or destroyed, only rearranged.” However, the equation the student wrote in item 5 was not balanced, implying a struggle with higher-level application of the law of conservation of matter in spite of prior factual identification. This alternative conception was observed across multiple students’ responses. While 106 out of the 182 students had either explicitly mentioned “law of conservation of mass” or suggested conservation of matter in their written answers to a science question, more complex items on the assessment yielded different results. On a related task that required the students to demonstrate their *understanding* of the law of conservation of mass—going beyond memorizing a physical principle—wherein they had to balance a chemical equation, only 41 out of the 182 correctly used the right coefficients in the equation. The inability of the students to enact their understanding of law of conservation of matter when it came to the integrated task highlights an affordance of using a well-designed assessment. During the professional development sessions prior to project implementation, the teachers assured us that balancing an equation (specifically the reversible equation for water forming and splitting) could be considered as prior knowledge for their students. However, our findings (such as those in Figs. 3, 4, and 5) suggest that within the context of integrated assessment tasks, we may uncover such overlooked misconceptions.

Drawing from Bloom’s Taxonomy (1956, revised version [Krathwohl 2002]), providing the Law of Conservation of Mass required only that students “remember,” the lowest level of cognitive demand. Alternative conceptions were most evident in the blended items, and we posit that this is due to the higher cognitive demand required in application of both learning domains. In the blended question, to correctly change the code for the breakdown of water to the breakdown of glucose, students needed to “analyze” (Level 4.0; Krathwohl 2002) the components of both the chemical equation for the breakdown of glucose and also the code provided in order to “create” (Level 6.0) new code for implementing the breakdown of glucose. This required representation of the correct reactants and products with the right molecule amounts as “interpret[ed]” (Level 2.1) from the chemical equation. It also required that the students interpreted the given code so that they knew how to use it as a guide for the new procedure. For example, they would need to “understand” (Level 2.0) that “hatch” creates new molecules and therefore should be used prior to introducing the products in the simulation. This reflects the changing nature of science education, with recent directives that science curricula incorporate core ideas, practices of scientific reasoning, and cross-cutting concepts (NRC 2012). As Pellegrino (2012) states, “We are moving beyond vague terms such as ‘know’ and ‘understand’ to more specific statements like analyze, compare, explain, argue, represent, predict, etc. in which the practices of science are wrapped around and integrated with core content (p.832).” Students’ struggles with the higher cognitive demand that are characteristic of such integrated tasks is evidenced by the assessment data presented here.



You want to modify this section of code to make your own model representing the breakdown of glucose. You know that the balanced chemical equation for the breakdown of glucose is as follows:

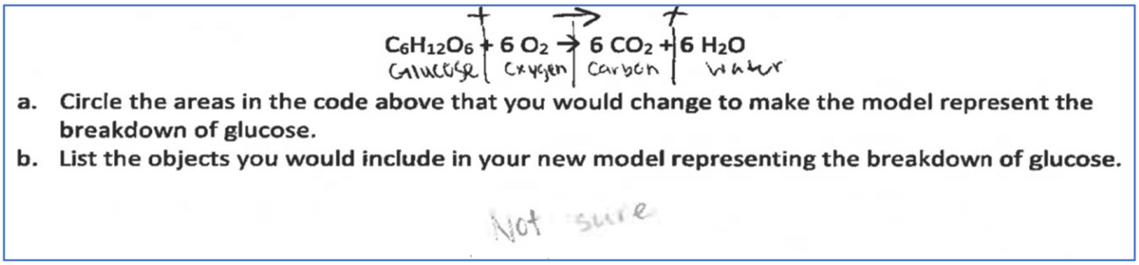
$C_6H_{12}O_6 + 6\,O_2 \rightarrow 6\,CO_2 + 6\,H_2O$ Change amount

a. Circle the areas in the code above that you would change to make the model represent the breakdown of glucose.

b. List the objects you would include in your new model representing the breakdown of glucose.

C, H, O

Fig. 4 S2 lists chemical elements to add to model



$C_6H_{12}O_6 + 6\ O_2 \rightarrow 6\ CO_2 + 6\ H_2O$

Glucose | Oxygen | Carbon | Water

a. Circle the areas in the code above that you would change to make the model represent the breakdown of glucose.

b. List the objects you would include in your new model representing the breakdown of glucose.

Not sure

Fig. 5 S3 recognizes the molecules in the equation

A last observation from S1's response was a mismatch in expected representation of learning according to the rubric and actual representational choices made by students. Given the complexity involved in representing atoms in our simulation, our NetLogo simulations had represented molecules. Our rubric too reflected our stance of focusing on the molecular representation of objects. Yet, students proposed representing atoms on the assessment. Thus, for 5b, the rubric specifies that students should provide the following objects for one point each: glucose ($C_6H_{12}O_6$), oxygen (O_2), carbon dioxide (CO_2), and water (H_2O). Student 1 suggested adding "C" and "O." Student 2, along with many other students, does something similar, suggested adding "C," "H," and "O" to the model (Fig. 4). The student did not receive points for this response, but each of these are atoms present in the molecules that compose the equation for the breakdown of glucose, which is perfectly correct.

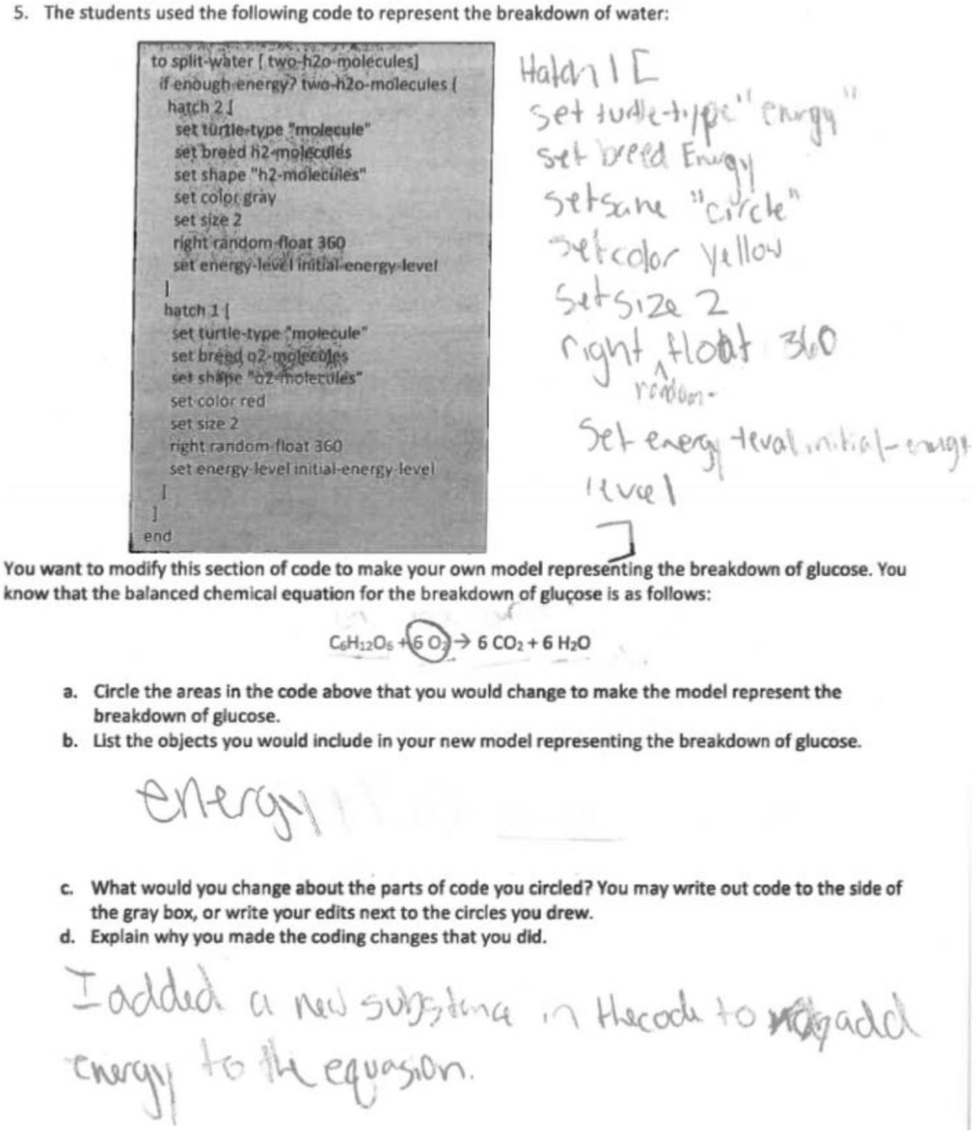
In contrast, S3 (Fig. 5) correctly identified several molecules by name in the chemical equation. Similar to S1 and S2, S3 had struggled with the pre-test. All students supplied varied forms of responses, including "IDK" and "I don't know" for questions 5a and 5b. In the post-test, we saw two different ways of responding. All three responses, to a greater extent, were correct. S1 and S2 demonstrated an understanding of atomic composition of chemical substances, although it remains unclear whether they understand which objects were the reactants and which were the products and how the reactants were transformed. In contrast, S3 appeared to grasp chemical formulas enough to translate the formula into common names. However, the three students' lack of vocabulary, particularly in comprehending "list of objects," appeared to hinder their performance in 5b.

On the whole, students' varied approaches to molecular versus atomic versus nomenclature-based representations show that students may choose different approaches to representing their learning. While the representations may not be incorrect, they also may not always align with the expectations of assessment designers or of the devised method for reliable scoring. Still, research has shown that the difference between atoms and molecules is difficult for students (Lee et al. 1993); thus, methods for assessing students' understanding in this area are important, and allowing multiple approaches may not always uncover such struggles.

Figure 6 presents an entirely different example of undetected learning—here in CT. Rather than changing the code to implement the breakdown of glucose as the task asked, S4 chose to add a new object, "energy," to the simulation. S4 did not interpret the product-reactant concepts represented in the equation of glucose breakdown (i.e., of glucose breaking down in the presence of oxygen resulting in the formation of water and carbon dioxide) but did demonstrate an understanding of energy being released subsequent to glucose breakdown—a concept that was previously covered in class. While our question aimed to prompt students to think about the chemical objects (glucose, oxygen, water, and carbon dioxide), S4 went beyond that to represent a science concept learned in class. The student drew on that knowledge to offer a new code-based model using appropriate coding structures for the implementation of a new object, "energy," including the designation of appropriate new properties (e.g., circle shape, yellow color, and with random movement). While we can only conjecture about S4's scientific reasoning, it is clear that the student was assessing and reasoning about the scientific model and the role of energy in glucose breakdown (as evident in the response to 5b and 5d), and drawing on that reasoning to think about abstractions and computational models. Models are abstracted representations of a certain aspect of a selected phenomenon created with a certain purpose (Giere, 2004). The abstraction could be at different levels and applied to any aspect of a phenomenon: it could be applied to the reactants and products in a chemical reaction, as we had done in our models and expected in the assessment, or could be applied to energy and its transformation, as S4 had done. The broadest kind of assessment would acknowledge this while evaluating a student's performance in assessing a model, a critical CT skill (Weintrop et al. 2015).

Discussion

This article reports on a larger project that explores the integration of CT with middle school science, with the aim of infusing CT practices into science instruction in a way that deepens students' understanding of chemistry concepts that are difficult to learn. Here, we focus on the challenges associated with and lessons learned from the design,



5. The students used the following code to represent the breakdown of water:

```
to split-water [ two-h2o-molecules]
  if enough-energy? two-h2o-molecules [
    hatch 2 [
      set turtle-type "molecule"
      set breed h2-molecules
      set shape "h2-molecules"
      set color gray
      set size 2
      right random-float 360
      set energy-level initial-energy-level
    ]
    hatch 1 [
      set turtle-type "molecule"
      set breed o2-molecules
      set shape "o2-molecules"
      set color red
      set size 2
      right random-float 360
      set energy-level initial-energy-level
    ]
  ]
end
```

```
Hatch 1 [
 set turtle-type "energy"
 set breed Energy
 setsane "circle"
 setcolor yellow
 Setsize 2
 right random float 360
 Set energy-leval initial-energy
 level
]
```

You want to modify this section of code to make your own model representing the breakdown of glucose. You know that the balanced chemical equation for the breakdown of glucose is as follows:

$C_6H_{12}O_6 + \boxed{6\ O_2} \rightarrow 6\ CO_2 + 6\ H_2O$

a. Circle the areas in the code above that you would change to make the model represent the breakdown of glucose.
b. List the objects you would include in your new model representing the breakdown of glucose.

energy

c. What would you change about the parts of code you circled? You may write out code to the side of the gray box, or write your edits next to the circles you drew.
d. Explain why you made the coding changes that you did.

I added a new substma in the code to ~~ma~~add energy to the equasion.

Fig. 6 S4 implements energy as a new object in the model

implementation, and evaluation of an integrated performance assessment. We describe our augmentation of quantitative analyses of a traditional rubric-based scoring approach with a second round of scoring that differentiated by discipline measured and qualitative analyses of students' responses to an open-ended task. We present a stage in the iterative development of an integrated CT and science intervention and associated assessment. In response to the analyses of data presented here and other data collected (e.g., video, student artifacts, teacher reflections), our team continues to develop curricular and assessment materials in tandem.

We report here on both partial success and complexity that arises as we learn more to fully integrate CT into core content. In particular, we argue that new forms of instruction require new forms of assessment. We agree with other work that advances integrated assessment and offer more depth to that work by pointing out how integration here changes the underlying epistemology of what is taught. Our work falls into the line of inquiry that investigates the benefit of open-ended assessments. Within this, we argue for multiple approaches to demonstrate learning, and in particular for scoring methods that are differentiated so that learning in each domain can be uncovered, including detailing the ways that the integration changes that epistemology. The benefit of this approach is not only a more sensitive instrument to provide evidence of learning but also an opportunity to identify alternative conceptions or gaps to inform instructors' next steps.

Epistomologically Pluralistic Assessments One of the affordances of learning science in the computational context is the opportunity to encounter concepts through multiple representations and, in turn, to approach problems through multiple avenues. Integrating computational systems affords multiple ways of knowing and doing. One of the creative benefits of computational modeling is that one can create representations of the same concept as seen through multiple lenses; we argue that assessments should reflect this affordance. Assessment design should not only support but encourage multiple ways of doing. Pragmatically, this can include open-ended items in which an objective is specified, yet a student may achieve this objective in multiple ways. In our context, students were asked to create or modify code to implement a scientifically accurate phenomenon. The student may receive particular concepts that must be modeled, but the task can be approached through multiple avenues. In this way, the assessment itself becomes a learning exercise affording the student with the opportunity to demonstrate his or her learning utilizing a rich array of acquired resources and skills. Such activities mirror real-world scientific practices and provide a rich learning environment (Järvelä 1995). This environment holds potential both for higher cognitive demand applications of learning and also for the identification of students' struggles with concepts.

Uncovering Alternative Conceptions Another benefit of the rich learning environment created by integrating CT with other domains is that it can reveal alternative conceptions. The nature of integration creates high-cognitive demand tasks, as each domain is applied to the other. We have seen that students have difficulty understanding the distinction between atomic and molecular representations and models. Normal science instruction moves often elide the differences, because it is so easy to move between them in language and chemical formulas contain both atomic and molecular representations. But in CT, abstraction means that at any given moment, we must choose either atoms or molecules and stay with them. Thus, solving science problems in the CT context puts a higher demand on the student. Our observations are that these tasks reveal whether a student is able to go beyond "remembering" the difference to actually applying it systematically to a more complex integrated task. In this way, such analyses can then also be powerful formative assessment tools (Basu et al. 2018), informing next steps for instruction.

Differentiated Scoring Since an integrated assessment actually mirrors the interdisciplinary nature of real-world tasks, it could be questioned whether it is important to know the specific domain in which a student struggled. We argue, a priori, that in the classroom context, domain-specific conclusions about student learning are important, for example to inform adjustments to future instruction. Therefore, if assessments consist of integrated tasks, we need a scoring method that is sensitive to the detection of learning in each domain *and* the interplay of domains to make conclusions about the nature of students' strengths and weaknesses. This argument extends a similar approach used for scoring project-based learning (Capraro and Corlu 2013; Petkov and Petkova 2006) to the domain of assessment. Validity lies in "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA and NCME 1999, p.9). Our use of a differentiated rubric increases validity by showing that students learned some aspects of one of the domains, but their performance on blended items was, at times, hindered by limited understanding of the other domain. Our analyses reveal that an integrated approach can result in significant learning gains. Yet, decoding the precise nature of the gain can be difficult. Looking at items from the point of view of differentiated scoring can add sensitivity to the way we assess learning. An important next step is to investigate approaches to the design of blended assessment tasks and accompanying rubrics that will allow us to detect that sensitivity more efficiently.

Conclusions

The current work struggles with the nature of this assessment, proposing that we integrate the assessment but differentiate the scoring. We seek to balance the need for efficient methods of administration and scoring with sensitivity to the complex nature of the learning we are trying to engender. Our message is both a proposal and a caution: First, as people attempt to create efficient approaches to assessment, it is crucial that they not overlook the most important benefits of the integrated approach—enhanced encounters with computing methods of expression tailored, in this case, for science. Designing assessments and the accompanying rubrics are iterative endeavors, similar to that of designing instructional modules. Our analysis suggests potential in using an open-ended assessment that supports and encourages multiple ways of doing. We recognize that the level of qualitative analysis conducted for this research cannot practically be undertaken by each classroom teacher. While qualitative evaluation of open-ended assessment responses is ideal, it is time consuming. A teacher or researcher may instead choose to seek opportunities to qualitatively evaluate students' progress on formative assessments or to conduct qualitative analysis on only a sample of student assessments, in order to inform next instructional moves. Second, the desire to use integrated assessment tasks in order to match instruction may lead to overlooked alternative conceptions or important gains in one domain or the other. Therefore, we see reason to include, in addition to integrated items, items assessing multiple facets from each domain that are not fully integrated. Our findings suggest that the changes to the instructional context as a result of integrating CT necessitate such an approach at least until we have a deeper understanding of the interplay of elements.

As multiple stakeholders explore potential routes for integrated CT assessment, such as automated scoring, we recommend consideration of the need for assessments that afford students the same opportunities to "think" and "do" in multiple ways. The interplay of integrated domains in an assessment context may benefit from specified identification of domains assessed by particular assessment tasks to produce a fuller picture of students' strengths and weaknesses in the various domains. Future rubric designs should attempt to allow for multiple expressions of learning while also linking criteria to the domain addressed. This is consistent with the call for systems of assessment (Grover 2015; Pellegrino 2012). All purposes cannot be served by one assessment.

Funding This study was funded by the National Science Foundation (grant number 1543022).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflicts of interest.

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educ Assess, 8*(3), 231–257. Retrieved from. https://doi.org/10.1207/S15326977EA0803_02.

AERA, A., & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Airasian, P. W., & Miranda, H. (2002). The role of assessment in the revised taxonomy. *Theory Pract, 41*(4), 249–254. Retrieved from. https://doi.org/10.1207/s15430421tip4104_8.

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *J Learn Sci, 13*(1), 1–14. https://doi.org/10.1207/s15327809jls1301_1.

Basu, S., Biswas, G., & Kinnebrew, J. S. (2017). Learner modeling for adaptive scaffolding in a computational thinking-based science learning environment. *User Model User-Adap Inter, 27*(1), 5–53.

Basu, S., McElhaney, K. W., Grover, S., Harris, C., & Biswas, G. (2018). *A principled approach to designing assessments that integrate science and computational thinking. In ICLS Proceedings*. London: ISLS.

Bell, P., Hoadley, C. M., & Linn, M. C. (2004). Design-based research in education. *Internet Environments for Science Education*, 73–85.

Bell, T., Witten, I. H., Fellows, M., Adams, R., McKenzie, J., Powell, M., & Jarman, S. (2015). CS unplugged: Computer science without a computer. Retrieved from https://csunplugged.org/en/

Biggs, J. (1996). Assessing learning quality: Reconciling institutional, staff and educational demands. *Assess Eval High Educ, 21*(1), 5–16. Retrieved from. https://doi.org/10.1080/0260293960210101.

Blikstein, P., & Wilensky, U. (2009). An atom is known by the company it keeps: A constructionist learning environment for materials science using agent-based modeling. *Int J Comput Math Learn, 14*(2), 81–119. Retrieved from. https://doi.org/10.1007/s10758-009-9148-8.

Box, G. E. (1976). Science and statistics. *J Am Stat Assoc, 71*(356), 791–799.

Box, G. E. (1979). All models are wrong, but some are useful. *Robustness in Statistics, 202*.

Brennan, K., & Resnick, M. (2012). New frameworks for studying and assessing the development of computational thinking. In *In Proceedings of the 2012 annual meeting of the American Educational Research Association*. Vancouver: Canada Retrieved from http://scratched.gse.harvard.edu/ct/files/AERA2012.pdf.

Buffum, P. S., Lobene, E. V., Frankosky, M. H., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2015). A practical guide to developing and validating computer science knowledge assessments with application to middle school. In *Proceedings of the 46th ACM technical symposium on computer science education* (pp. 622–627). New York, NY: ACM. Retrieved from. https://doi.org/10.1145/2676723.2677295.

Capraro, R. M., & Corlu, M. S. (2013). Changing views on assessment for STEM project-based learning. In *STEM project-based learning* (pp. 109–118). Rotterdam, NL: SensePublishers.

Dickes, A., & Sengupta, P. (2013). Learning natural selection in 4th grade with multi-agent-based computational models. *Res Sci Educ, 43*(3), 921–953. Retrieved from. https://doi.org/10.1007/s11165-012-9293-2.

diSessa, A. (2000). *Changing Minds*. Cambridge, MA: The MIT Press.

Gane, B. D., Zaidi, S. Z., & Pellegrino, J. W. (2018). Measuring what matters: Using technology to assess multidimensional learning. *Eur J Educ, 53*, 176–187.

Gautam, A., Wall Bortz, W., & Tatar, D. (2017). *Case for integrating computational thinking and science in a low-resource setting. In Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*. Pakistan: Lahore.

Giere, R. N. (2004). How Models Are Used to Represent Reality. Philosophy of Science, 71(5), 742–752. https://doi.org/10.1086/425063.

Goldstone, R. L., & Wilensky, U. (2008). Promoting transfer by grounding complex systems principles. *J Learn Sci, 17*(4), 465–516. Retrieved from. https://doi.org/10.1080/10508400802394898.

Grover, S. (2015). *Systems of assessments for deeper learning of computational thinking in K - 12, 10*. Chicago: Presented at the American Educational Research Association.

Grover, S., & Pea, R. (2013). Computational Thinking in K–12: A Review of the State of the Field. *Educational Researcher, 42*(1), 38–43. https://doi.org/10.3102/0013189X12463051.

Grover, S., & Pea, R. (2018). Computational thinking: A competency whose time has come. In S. Sentance, S. Carsten, & E. Barendsen (Eds.), *Computer science education: Perspectives on teaching and learning in school* (pp. 19–38). London, UK: Bloomsbury.

Herrmann-Abell, C. F., & DeBoer, G. E. (2011). Using distractor-driven standards-based multiple-choice assessments and Rasch modeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items. *Chemistry Education Research and Practice, 12*(2), 184–192.

Herrmann-Abell, C. F., Koppal, M., & Roseman, J. E. (2016). Toward high school biology: Helping middle school students understand chemical reactions and conservation of mass in nonliving and living systems. *CBE-Life Sciences Education, 15*(4), 1–21. Retrieved from. https://doi.org/10.1187/cbe.16-03-0112.

Hickey, D. T., & Zuiker, S. J. (2012). Multilevel aAssessment for dDiscourse, uUnderstanding, and aAchievement. *J Learn SciJournal of the Learning Sciences, 21*(4), 522–582. https://doi.org/10.1080/10508406.2011.652320.

Hoadley, C. M. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educ Res, 32*(1), 5–8.

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. *Comput Educ, 82*, 263–279. Retrieved from. https://doi.org/10.1016/j.compedu.2014.11.022.

Jacobson, M. J., & Wilensky, U. (2006). Complex systems in education: Scientific and educational importance and implications for the learning sciences. *J Learn Sci, 15*(1), 11–34. Retrieved from. https://doi.org/10.1207/s15327809jls1501_4.

Järvelä, S. (1995). The cognitive apprenticeship model in a technologically rich learning environment: Interpreting the learning interaction. *Learn Instr, 5*(3), 237–259. Retrieved from. https://doi.org/10.1016/0959-4752(95)00007-P.

Kaput, J., & Schorr, R. (2008). Changing representational infrastructures changes most everything: The case of SimCalc, algebra, and calculus. In K. Heid & G. W. Blume (Eds.), *Research on technology and the teaching and learning of mathematics: Case and perspectives* (Vol. 2, pp. 211–253). Charlotte, NC: Information Age Publishing.

Koh, K. H., Basawapatna, A., Bennett, V., & Repenning, A. (2010). Towards the automatic recognition of computational thinking for adaptive visual language learning. In *2010 IEEE Symposium on Visual Languages and Human-Centric Computing* (pp. 59–66). Retrieved from**m**). https://doi.org/10.1109/VLHCC.2010.17.

Koh, K. H., Basawapatna, A., Nickerson, H., & Repenning, A. (2014). Real time assessment of computational thinking. In *2014 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (pp. 49–52). Retrieved from). https://doi.org/10.1109/VLHCC.2014.6883021.

Krajcik, J., Blumenfeld, P. C., Marx, R. W., Bass, K. M., Fredricks, J., & Soloway, E. (1998). Inquiry in project-based science classrooms: Initial attempts by middle school students. *J Learn Sci, 7*(3–4), 313–350.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Pract, 41*(4), 212–218.

Lead States, N. G. S. S. (2003). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.

Lee, O., Eichinger, D. C., Anderson, C. W., Berkheimer, G. D., & Blakeslee, T. D. (1993). Changing middle school students' conceptions of matter and molecules. *J Res Sci Teach, 30*(3), 249–270. Retrieved from. https://doi.org/10.1002/tea.3660300304.

Lee, I., Martin, F., Denner, J., Coulter, B., Allan, W., Erickson, J., et al. (2011). Computational thinking for youth in practice. *ACM Inroads, 2*(1), 32–37. Retrieved from. https://doi.org/10.1145/1929887.1929902.

Lye, S. Y., & Koh, J. H. L. (2014). Review on teaching and learning of computational thinking through programming: What is next for K - 12? *Comput Hum Behav, 41*, 51–61. Retrieved from. https://doi.org/10.1016/j.chb.2014.09.012.

MacQueen, K. M., McLellan, E., Kay, K., & Milstein, B. (1998). Codebook development for team-based qualitative analysis, codebook development for team-based qualitative analysis. *CAM Journal, 10*(2), 31–36. Retrieved from. https://doi.org/10.1177/1525822X980100020301.

Marzano, R. J., Pickering, D., & McTighe, J. (1993). *Assessing student outcomes: Performance assessment using the dimensions of learning model*. Alexandria, VA: Association for Supervision and Curriculum Development Retrieved from http://eric.ed.gov/?id=ED461665.

Meerbaum-Salant, O., Armoni, M., & Ben-Ari, M. (2013). Learning computer science concepts with scratch. *Comput Sci Educ, 23*(3), 239–264.

Miller, D. M., & Linn, R. L. (2000). Validation of performance-based assessments. *Appl Psychol Meas, 24*(4), 367–378. Retrieved from. https://doi.org/10.1177/01466210022031813.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educ Meas Issues Pract, 25*(4), 6–20. Retrieved from. https://doi.org/10.1111/j.1745-3992.2006.00075.x.

Moreno-León, J., & Robles, G. (2015). Dr. scratch: A web tool to automatically evaluate scratch projects. In *Proceedings of the workshop in primary and secondary computing education* (pp. 132–133). New York, NY: ACM.

Moskal, B. (2000). Recommendations for developing classroom performance assessments and scoring rubrics. *Pract Assess Res Eval, 8*(14), 1–8.

National Research Council. (2011). *Report of a workshop on the pedagogical aspects of computational thinking*. Washington, DC: National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

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

Pellegrino, J. W. (2012). Assessment of science learning: Living in interesting times. *J Res Sci Teach, 49*(6), 831–841. Retrieved from. https://doi.org/10.1002/tea.21032.

Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science, 340*(6130), 320–323. Retrieved from. https://doi.org/10.1126/science.1232065.

Pellegrino, J. W., & Hilton, M. L. (2013). Committee on defining deeper learning and 21st century skills, Center for Education, Division on Behavioral and Social Sciences and Education, and National Research Council. In *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington DC: National Academies Press.

Perković, L., Settle, A., Hwang, S., & Jones, J. (2010). A framework for computational thinking across the curriculum. In *Proceedings of the fifteenth annual conference on innovation and technology in computer science education* (pp. 123–127). New York, NY: ACM.

Petkov, D., & Petkova, O. (2006). Development of scoring rubrics for IS projects as an assessment tool. *Issues in Informing Science and Information Technology, 3*, 499–510.

Qualls, J. A., & Sherrell, L. B. (2010). Why computational thinking should be integrated into the curriculum. *Journal of Computer Sciences in Colleges, 25*(5), 66–71.

Reed, D. A., Bajcsy, R., Fernandez, M. A., Griffiths, J.-M., Mott, R. D., Dongarra, J., et al. (2005). *Computational science: Ensuring America's competitiveness*. Arlington, VA: President's Information Technology Advisory Committee Retrieved from http://www.dtic.mil/docs/citations/ADA462840.

Román-González, M. (2015). *Computational thinking*. Test: Design Guidelines and Content Validation. Retrieved from. https://doi.org/10.13140/RG.2.1.4203.4329.

Sampson, V., & Grooms, J. (2008). Science as argument-driven inquiry: The impact on students' conceptions of the nature of scientific inquiry. In *Annual International Conference of the National Association of Research in Science Teaching*. MD: Baltimore.

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., et al. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *J Res Sci Teach, 46*(6), 632–654.

Sendur, G., Toprak, M., & Pekmez, E. S. (2010). *Analyzing of students' misconceptions about chemical equilibrium. International Conference on New Trends in Education and Implications*. Turkey: Antalya.

Sengupta, P., Kinnebrew, J. S., Basu, S., Biswas, G., & Clark, D. (2013). Integrating computational thinking with K - 12 science education using agent-based computation: A theoretical framework. *Educ Inf Technol, 18*(2), 351–380. Retrieved from. https://doi.org/10.1007/s10639-012-9240-x.

Settle, A., & Perkovic, L. (2010). Computational thinking across the curriculum: A conceptual framework. In *Technical reports* Retrieved from http://via.library.depaul.edu/tr/13.

Settle, A., Franke, B., Hansen, R., Spaltro, F., Jurisson, C., Rennert-May, C., & Wildeman, B. (2012). Infusing computational thinking into the middle- and high-school curriculum. In *Proceedings of the 17th ACM annual conference on innovation and Technology in*

Computer Science Education (pp. 22–27). New York, NY, USA: ACM. Retrieved from. https://doi.org/10.1145/2325296.2325306.

Sherman, M., & Martin, F. (2015). The assessment of mobile computational thinking. *Journal of Computing Sciences in Colleges, 30*(6), 53–59.

Simon, H. A. (1955). A behavioral model of rational choice. *Q J Econ, 69*, 99–118.

Stavy, R. (1991). Children's ideas about matter. *Sch Sci Math, 91*(6), 240–244.

Texas Education Agency. (2013). *Texas essential knowledge and skills for science*. Austin, TX: Texas Education Agency Retrieved from http://ritter.tea.state.tx.us/rules/tac/chapter112/ch112b.html.

Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data, a general inductive approach for analyzing qualitative evaluation data. Am J Eval, 27(2), 237–246. Retrieved from https://doi.org/10.1177/1098214005283748

Turkle, S., & Papert, S. (1992). Epistemological pluralism and the revaluation of the concrete. *Journal of Mathematical Behavior, 11*(1), 3–33.

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

Wall Bortz, W., Gautam, A., Tatar, D., Rivale, S., & Lipscomb, K. (2019). The availability of pedagogical responses and the integration of computational thinking. In M. Reardon & J. Leonard (Eds.), *Integrating digital technology in education: School-university-community collaboration*. Charlotte, NC: Information Age Publishing.

Weintrop, D., Beheshti, E., Horn, M. S., Orton, K., Trouille, L., Jona, K., & Wilensky, U. (2014). Interactive assessment tools for computational thinking in high school STEM classrooms. In *Intelligent Technologies for Interactive Entertainment* (pp. 22–25). Chicago: IL.

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2015). Defining computational thinking for mathematics and science classrooms. *J Sci Educ Technol, 25*(1), 127–147. Retrieved from. https://doi.org/10.1007/s10956-015-9581-5.

Werner, L., Denner, J., Campe, S., & Kawamoto, D. C. (2012). The fairy performance assessment: Measuring computational thinking in middle school. In *Proceedings of the 43rd ACM technical symposium on computer science education* (pp. 215–220). New York, NY: ACM.

Wiggins, G. (1990). The case for authentic assessment. *Pract Assess Res Eval, 2*(2).

Wilensky, U. (1999). *NetLogo*. Evanston, IL: Northwestern University, Center for Connected Learning and Computer-Based Modeling Retrieved from http://ccl.northwestern.edu/netlogo/.

Wilensky, U., & Reisman, K. (2006). Thinking like a wolf, a sheep, or a firefly: Learning biology through constructing and testing computational theories-an embodied modeling approach. *Cogn Instr, 24*(2), 171–209.

Wilensky, U., & Stroup, W. (1999). Learning through participatory simulations: Network-based design for systems learning in classrooms. In *Proceedings of the 1999 Conference on computer support for collaborative learning*. Palo Alto, California: International Society of the Learning Sciences.

Wing, J. M. (2006). Computational thinking. *Commun ACM, 49*(3), 33–35.

Yadav, A., Mayfield, C., Zhou, N., Hambrusch, S., & Korb, J. T. (2014). Computational thinking in elementary and secondary teacher education. *ACM Trans Comput Educ, 14*(1), 5. Retrieved from. https://doi.org/10.1145/2576872.

Yaşar, O. (2018). A new perspective on computational thinking. *Commun ACM, 61*(7), 33–39.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.